The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Information Technology
Master of Information Technology

الجامعـــــــة الإســـلاميـة – غـزة
شئون البحث العلمي والدراسات العليا
كـليـــة تكنولوجيـــــا المعلومــــــات
ماجستير تكنولـــوجيا المعلومــــــات

Semi-Automatic Method for Infoboxes Extraction for Arabic Wikipedia Articles

طريقة شبه آلية لاستخلاص معلومات مختصرة لمقالات ويكيبيديا العربية

Saleem M. Shublaq

Supervised by

Rawia F. Awadallah

Assistant Professor

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Information Technology

December/2016

إقــــــــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Semi-Automatic Method for Infoboxes Extraction for Arabic Wikipedia Articles

طريقة شبه آلية لاستخلاص معلومات مختصرة لمقالات ويكيبيديا

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	**سليم محمد شبلاق**	اسم الطالب:
Signature:		التوقيع:
Date:	2017 - 03 - 04	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية – غزة
The Islamic University - Gaza

الجامعة الإسلامية غزة
The Islamic university Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا   هاتف داخلي: 1150

الرقم ج س غ/35/ Ref
التاريخ 2016/12/29 Date

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ سليم محمد سليم شبلاق لنيل درجة الماجستير في كلية *تكنولوجيا المعلومات* برنامج تكنولوجيا المعلومات وموضوعها:

طريقة شبه آلية لاستخلاص معلومات مختصرة من مقالات ويكيبيديا

Semi-Automatic Method for Infoboxes Extraction for Arabic Wikipedia Articles

وبعد المناقشة التي تمت اليوم الأربعاء 29 ربيع أول 1437هـ، الموافق 2016/12/28م الساعة الواحدة ظهراً، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. رواية فوزي عوض الله	مشرفاً و رئيساً	....
أ.د. علاء مصطفى الهليس	مناقشاً داخلياً	....
د. يوسف نبيل أبو شعبان	مناقشاً خارجياً	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق ،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

شئون البحث العلمي والدراسات العليا
الجامعة الإسلامية - غزة
The Islamic University - Gaza
Research & Graduate Affairs

Abstract

Arabic Language is one of the popular languages over the world. There is 5% of people over the world speak Arabic. However, it suffers from a low percentage of content over the internet. Wikipedia is a very well-known multilingual, web-based, free-content encyclopedia project supported by the Wikimedia Foundation and based on a model of openly editable content. It is one of the greatest repositories of human knowledge ever constructed, and has high ranks in Google that makes its pages often pop up in search results. Arabic Wikipedia, which is part of Wikipedia website, lacks valuable content compared to Wikipedia content for other languages. Besides, many of existing articles are stub pages containing only one or few sentences of text that is too short to provide encyclopedic coverage of a subject. Some researchers worked on increasing and enriching the content of Wikipedia, but most of these efforts focused on developing methods that process text in other languages rather than Arabic.

This research aims at boosting online Arabic content. In particular, it aims to boost the editing process in Arabic Wikipedia. Our main objective is to develop method for suggesting contents for Arabic Wikipedia articles either to enrich the contents of existing stub pages or to generate new ones that contain infobox. The proposed methods build on existing methods in Information Retrieval, Question Answering, and Text Mining in order to extract key information from relevant documents on the web. The automatically generated contents and the different resources from which these contents are extracted will be available for Wikipedia editors for revision and proofreading before adding them to Wikipedia. In this research, we focus on enriching the Infobox which is a summary of some unifying parameters at the top left/right corner of an article. We developed four main algorithms to extract) birth, death)locations, (birth, death) dates and full name of entity. We have conducted many experiments to evaluate our methods on articles about named entities in the political domain. Our results achieved an overall accuracy of 80.3%.

Keywords: Wikipedia, Infobox, Information Retrieval

الملخص

تعتبر اللغة العربية واحدة من أكثر اللغات إنتشارا في العالم. هناك ما يقارب 5% من سكان العالم يتحدثون اللغة العربية، على الرغم من ذلك فإن اللغة العربية تعاني من ضعف نسبة المحتوى العربي على الانترنت. ويكيبيديا هو موقع ويب يحتوي على موسوعة من المقالات متعددة اللغات ومدعوم من مؤسسة ويكيميديا. هو واحد من أكبر قواعد البيانات التي تحتوي على المعرفة في مختلف المجالات حتى اللحظة بالإضافة لانه من المواقع التي تستحوذ على ترتيب مترفع في تصنيف جوجل حيث أنه يظهر دائما في أولى نتائج البحث.

النسخة العربية من ويكيبيديا هي جزء من موقع ويكيبيديا ولكنه يعتبر ذو محتوى ضعيف مقارنة بمحتوى اللغات الأخرى. هناك عدد كبير من المقالات الموجودة في النسخة العربية من ويكيبيديا تعتبر مقالات (بذرة) وهي مقالات تحتوي على معلومات قليلة جدا عن موضوع المقال. بعض الباحثين عملوا على زيادة وإثراء المحتوى الخاص ب ويكيبيديا، لكن أغلب الجهود مركزة على اللغات الأخرى غير اللغة العربية.

هذا البحث يهدف لزيادة المحتوى الرقمي العربي بالإضافة لتحسين عملية التعديل الخاصة بالمقالات العربية. يركز الهدف الأساسي للبحث على تطوير طريقة لإقتراح محتوى لمقالات ويبيكيديا العربية لإثراء المحتوى الخاص بمقالات (بذرة) العربية أو إنشاء مقالات جديدة. تأتي هذه الطريقة المقترحة بالاعتماد على طرق تابعة لمجال إستخراج المعلومات، الأسئلة المجابة والتنقيب عن النصوص لاستخلاص المعلومات من المستندات والوثائق الموجودة على الويب. المعلومات التي يتم استخراجها ستكون متاحة للمحررين للمراجعة والتعديل قبل اضافتها على ويكيبيديا. في هذا البحث تم التركيز على إثراء صندوق المعلومات والذي يعتبر ملخص للمقالات ويتم وضعه في الزاوية اليمنى / اليسرى للمقال. قمنا بتطوير أربعة خوارزميات رئيسية تهدف لاستخلاص مكان (ميلاد، وفاة) وتاريخ (ميلاد، وفاة) والإسم الكامل للشخصية التي يتم البحث عنها بالإضافة لعمل خوارزمية التصويت التي تعمل على كل خوارزمية لزيادة دقة النتيجة المرشحة للظهور للمستخدم. قمنا بعمل العديد من التجارب لقياس أداء الخوارزميات وإقتصرت التجارب على المجال السياسي، وحققت هذه التجارب دقة نسبتها 80.3 %.

Dedication

أهدي هذا النجاح الى صانعيه الحقيقيين، إلى من لم يدخروا جهداً حتى يتكلل هذا العمل
بالنجاح، أهديه لمن وقفوا بجانبي وقدموا لي كل ما أحتاج من الدعم حتى وصلت لختام هذا
العمل، أهديه لأبي ... أمي ... وعائلتي

وإلى من شاركوني هذا النجاح

الصديق الدكتور المهندس محمد شراب الذي لم يدخر جهدا ليقدم لي الدعم والنصيحة
والتشجيع والمساعدة في كل وقت

صديقي العزيز أحمد عابد الذي يشاركني كل نجاحاتي ويدعمني دائما

وأتمنى من الله عز وجل أن يكون هذا العمل بداية الطريق لمزيد من التقدم والنجاح

والحمد لله رب العالمين الذي وفقني لهذا

Acknowledgment

I would like to express my deepest gratitude and appreciation to my advisor Dr. Rawia Awadallah, for her hard work and guidance throughout this entire thesis process and for believing in my abilities. I have learned so much, and without her, this would not have been possible. Thank you so much for the great experience.

Also I am extremely thankful to all members of the Information Technology Faculty staff at the Islamic University of Gaza, for their assistance.

Much gratitude is given to my family, for their never-ending love and support in all my efforts, and for giving me the foundation to be who I am. Thank you all my friends for support.

Saleem M. Shubaq

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ATKS	Arabic ToolKit Server
GATE	General Architecture for Text Engineering
GPE	Global Political Economy
NER	Named Entity Recognition
SWAA	Wikipedia Articles in Arabic Language

Chapter 1

Introduction

Chapter 1

Introduction

Internet became a cultural and educational encyclopedia for all areas. It deals with a large number of the world's languages. Nowadays, it is considered as a bowl to publish books through digital libraries. It is currently used as a media tool for the definition of peoples and states, since this network became the easiest and fastest way to get the information. One of the top most popular websites is Wikipedia. It is an online multilingual encyclopedia which contains many articles. The importance of Wikipedia comes from the huge human knowledge about different topics and in different domains such as; sport, politics, culture, etc. In addition, its pages are often on the top of the search results of most search engines such as Google. There are 280 languages available on Wikipedia. Arabic language is among them. It is the most spoken language among the Semitic languages group and one of the most widespread languages over the world. Arabic language has a special importance for Muslims because it is a religion related language (i.e. language of Quran, prayers, and worship acts) (Wikipedia).

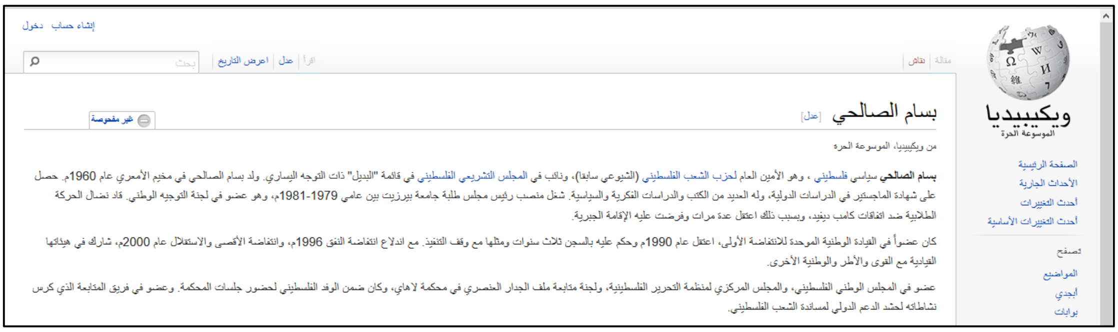
Although Arabic language is very popular, it is noted that there is a certain weakness of its online contents. The statistics showed that the proportion of digital content for Arabic Language is scarce and only 0.8 % of this content is in the top 10 million websites over the world (Wikipedia). Besides, Arabic language has only about 450,000 articles on Wikipedia, although it is spoken by about 422 million persons, while, Norwegian language is spoken by 4.6 million people and has about 300,000 articles in Wikipedia (Wikipedia). Moreover, most existing Arabic articles are too short to provide encyclopedic coverage of a subject. Such articles that belong to 'stub' Wikipedia category with few sentences are known as 'stub' articles in Wikipedia notions. An example for a stub article is shown in Figure 1. Furthermore, many articles in Arabic Wikipedia lack the Infoboxes on the right/left-top corners of the pages which are usually used to summarize some unifying parameters. For instance, every politician has a name, date of birth, birthplace, nationality, and field of study. An example for an Infobox is shown in Figure 2. This weakness in Arabic Wikipedia can be attributed to slowness in its editing process. The lack of relevant resources is one of the problems that faces the editors, and as a result affects the growth average of Wikipedia content.

The proposed research came from the Arabic language necessity. It is obvious that Arabic language contents are not sufficient compared to other languages contents over the Internet. Based on recent statistics, it is showed that the percentage of growth in Arabic language content on the internet is 3% (Wikipedia). This growth is distributed over many websites. Most of these websites receive low ranks by famous search engines such as Google and thus they do not appear among the top returned results. On the other hand, Wikipedia articles are highly ranked by most search engines such as Google. But Arabic Wikipedia content on the Web is still lacking in spite of the increased in number of Internet users in the Arab world. Besides, most Arab Internet users look for Arabic content. Therefore, increasing the content of Arabic Wikipedia, makes more free knowledge accessible for many users. In addition, developing method that support the editing process in Arabic Wikipedia, accelerates the growth rate of the contents, and empowers and encourage a global volunteer community to develop the world's knowledge and to make it available to everyone for free, for any purpose.

Many research works are using techniques from Information Retrieval (IR), Question Answering (QA), Text Mining (TM), Information Extraction, and Named Entity Recognition NER for increasing and enriching the contents of Wikipedia in different languages, but there is a limitation in the amount of research work devoted to Arabic Wikipedia. Most Arabic language articles are too short to provide encyclopedic coverage of a subject. Furthermore, many articles in Arabic Wikipedia lack Infoboxes which are formatted tables usually appear at the top left/right corners of pages. Such an Infobox provides a summary as structured information of some unifying parameters (e.g. name, birthdate, age, etc.) about the subject of an article (see Figure 2 for an example).

There are some challenges facing the editing process which affect the average growth of Arabic Wikipedia. Among these challenges are the lack of online textual resources and structured information in Arabic language, the lack of financial resources for editors, and the limitations in time.

Developing method for partially automating the editing process in order to overcome these challenges, would definitely boost the whole process and helps the editors in their task. This method might for example provide resources and some structured information relevant to a particular subject of an article. In this research, we study how we can help in enriching the contents of Wikipedia Articles in Arabic language (WAA) in an automatic way. In other words, we ask one basic question: can we develop an automated method for suggesting contents for such articles? More specifically, can we develop a method that suggests infoboxes (with entities such as Full Name, Birth and Death Location and, Birth and Death Date) for WAA for persons (e.g. بسام الصالحي)?



بسام الصالحي

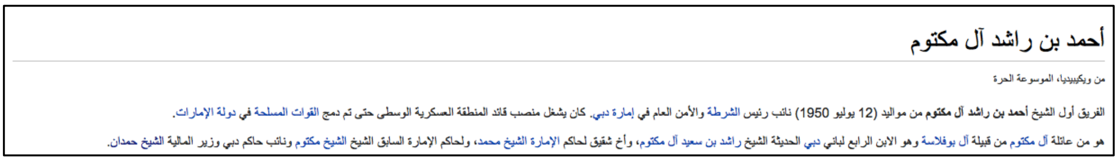
من ويكيبيديا، الموسوعة الحره

بسام الصالحي سياسي فلسطيني ، وهو الأمين العام لحزب الشعب الفلسطيني (الشيوعي سابقا)، ونائب في المجلس التشريعي الفلسطيني في قائمة "البديل" ذات التوجه اليساري. ولد بسام الصالحي في مخيم الأمعري عام 1960م. حصل على شهادة الماجستير في الدراسات الدولية، وله العديد من الكتب والدراسات الفكرية والسياسية. شغل منصب رئيس مجلس طلبة جامعة بيرزيت بين عامي 1979-1981م، وهو عضو في لجنة التوجيه الوطني. قاد نضال الحركة الطلابية ضد اتفاقات كامب ديفيد، وبسبب ذلك اعتقل عدة مرات وفرضت عليه الإقامة الجبرية.

كان عضواً في القيادة الوطنية الموحدة للانتفاضة الأولى، اعتقل عام 1990م وحكم عليه بالسجن ثلاث سنوات ومثلها مع وقف التنفيذ. مع اندلاع انتفاضة النفق 1996م، وانتفاضة الأقصى والاستقلال عام 2000م، شارك في هيئاتها القيادية مع القوى والأطر والوطنية الأخرى.

عضو في المجلس الوطني الفلسطيني، والمجلس المركزي لمنظمة التحرير الفلسطينية، ولجنة متابعة ملف الجدار العنصري في محكمة لاهاي، وكان ضمن الوفد الفلسطيني لحضور جلسات المحكمة. وعضو في فريق المتابعة الذي كرس نشاطاته لحشد الدعم الدولي لمساندة الشعب الفلسطيني.

Figure (1.1 a): Examples for Stub Articles



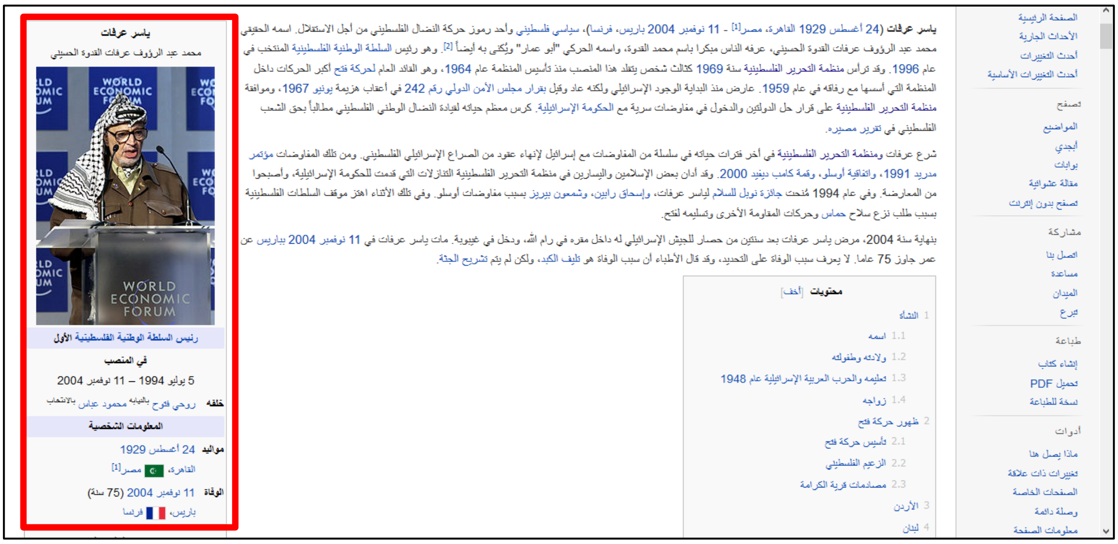
أحمد بن راشد آل مكتوم

من ويكيبيديا، الموسوعة الحرة

الفريق أول الشيخ **أحمد بن راشد آل مكتوم** من مواليد (12 يوليو 1950) نائب رئيس الشرطة والأمن العام في إمارة دبي. كان يشغل منصب قائد المنطقة العسكرية الوسطى حتى تم دمج القوات المسلحة في دولة الإمارات.

هو من عائلة آل مكتوم من قبيلة آل بوفلاسة وهو الابن الرابع لباني دبي الحديثة الشيخ راشد بن سعيد آل مكتوم، وأخ شقيق لحاكم الإمارة الشيخ محمد، ولحاكم الإمارة السابق الشيخ الشيخ مكتوم ونائب حاكم دبي وزير المالية الشيخ حمدان.

Figure (1.1 b): Examples for Stub Articles



ياسر عرفات (24 أغسطس 1929 القاهرة، مصر[1] - 11 نوفمبر 2004 باريس، فرنسا)، سياسي فلسطيني وأحد رموز حركة النضال الفلسطيني من أجل الاستقلال. اسمه الحقيقي محمد عبد الرؤوف عرفات القدوة الحسيني، عرفه الناس مبكرا باسم محمد القدوة، واسمه الحركي "أبو عمار" ويُكنى به أيضاً.[2] وهو رئيس السلطة الوطنية الفلسطينية المنتخب في عام 1996. وقد ترأس منظمة التحرير الفلسطينية سنة 1969 كثالث شخص يتقلد هذا المنصب منذ تأسيس المنظمة عام 1964، وهو القائد العام لحركة فتح أكبر الحركات داخل المنظمة التي أسسها مع رفاقه في عام 1959. عارض منذ البداية الوجود الإسرائيلي ولكنه عاد وقبل بقرار مجلس الأمن الدولي رقم 242 في أعقاب هزيمة يونيو 1967، وموافقة منظمة التحرير الفلسطينية على قرار حل الدولتين والدخول في مفاوضات سرية مع الحكومة الإسرائيلية. كرس معظم حياته لقيادة النضال الوطني الفلسطيني مطالباً بحق الشعب الفلسطيني في تقرير مصيره.

شرع عرفات ومنظمة التحرير الفلسطينية في أخر فترات حياته في سلسلة من المفاوضات مع إسرائيل لإنهاء عقود من الصراع الإسرائيلي الفلسطيني. ومن تلك المفاوضات مؤتمر مدريد 1991، واتفاقية أوسلو، وقمة كامب ديفيد 2000. وقد أدان بعض الإسلامين واليسارين في منظمة التحرير الفلسطينية التنازلات التي قدمت للحكومة الإسرائيلية، وأصبحوا من المعارضة. وفي عام 1994 مُنحت جائزة نوبل للسلام لياسر عرفات، وإسحاق رابين، وشمعون بيريز بسبب مفاوضات أوسلو. وفي تلك الأثناء اهتز موقف السلطات الفلسطينية بسبب طلب نزع سلاح حماس وحركات المقاومة الأخرى وتسليمه لفتح.

بنهاية سنة 2004، مرض ياسر عرفات بعد سنتين من حصار للجيش الإسرائيلي له داخل مقره في رام الله، ودخل في غيبوبة. مات ياسر عرفات في 11 نوفمبر 2004 بباريس عن عمر جاوز 75 عاما. لا يعرف سبب الوفاة على التحديد، وقد قال الأطباء أن سبب الوفاة هو تليف الكبد، ولكن لم يتم تشريح الجثة.

ياسر عرفات

محمد عبد الرؤوف عرفات القدوة الحسيني

رئيس السلطة الوطنية الفلسطينية الأول

في المنصب

5 يوليو 1994 – 11 نوفمبر 2004

خلفه	روحي فتوح بالنيابه محمود عباس بالانتخاب

المعلومات الشخصية

مواليد	24 أغسطس 1929 القاهرة، مصر[1]
الوفاة	11 نوفمبر 2004 (75 سنة) باريس، فرنسا

محتويات [أخف]

1 النشأة
1.1 اسمه
1.2 ولادته وطفولته
1.3 تعليمه والحرب العربية الإسرائيلية عام 1948
1.4 زواجه
2 ظهور حركة فتح
2.1 تأسيس حركة فتح
2.2 الزعيم الفلسطيني
2.3 مصادمات قرية الكرامة
3 الأردن
4 لبنان

Figure (1.2): An example for an Infobox of an article in Wikipedia (red box on the left)

1.1 Background and Context

The research that we describe in this thesis is related to many fields:

1.1.1 Information Retrieval

Information Retrieval (IR) is a step to access information resources that are relevant to information needed from a collection of information material (Saste & Patil, 2014). Information retrieval uses query that does not always return a single row of information. It returns many rows of information from database. The returned information has several forms such as text documents, images, and videos. Information Retrieval Field faces some challenges (Voniatis), some of them are:

a) Discovery

For any search engine, it is hard to find content if the web links are not known previously. Usually, some pages change their web address by updating URL, formatting, and content.

b) Storage

When search engines note that there is a change in web pages' content or there is a new content, it starts storing a copy of these websites pages. There are more than trillions of web pages on the World Wide Web. Therefore, it is not an easy task to store the huge number of content.

c) Extraction

We know that search engine provide users the searching service, sometimes search engine need to extract a specific elements and attribute from the web page, this process is not easy to do because of site architecture and content structure.

d) Modeling

This step is to transform the unstructured data into information. It uses techniques such as machine learning to detect the patterns for helping search engines.

1.1. 2 Information Extraction

Information Extraction (IE) is a sub task of information retrieval and it focuses on extracting structured information from unstructured one. This is based on natural language processing. IE has a sub tasks such as: named entity recognition, relationship extraction, and terminology extraction (Tari et al., 2012).

a) Open information extraction

Typically, Information Extraction (IE) systems learn an extractor for each target relation This approach to IE does not scale to corpora where the number of target relations is very large, or where the target relations cannot be specified in advance Open IE solves this problem by identifying relation phrases—phrases that denote relations in English sentences.

Open IE systems have achieved a notable measure of success on massive, open-domain corpora drawn from the Web, Wikipedia, and elsewhere (Fader, etal., 2011).

b) Arabic Word Net

In recent years, a number of WordNet building efforts have been initiated and carried out within a common framework for lexical representation and are becoming increasingly important resources for a wide range of Natural Language Processing applications (Elkateb et al., 2006).

The writing system of Arabic has twenty-five consonants and three long vowels that are written from right to left and take different shapes according to their position in the word. In addition to the long vowels, Arabic has short vowels Short vowels are not part of the alphabet but rather are written as vowel diacritics above or under a consonant to give it its desired sound and hence give a word a desired meaning.

Texts without vowels are considered to be more appropriate by the Arabic-speaking community since this is the usual form of everyday written and printed materials (books, magazines, newspapers, letters, etc.) and this difference make a challenges (Rodríguez, et al., 2008) (Elkateb et al., 2006).

Arabic WordNet is a recently initiated project that focus on building a lexical resource for Modern Standard Arabic based on the widely used Princeton WordNet for English Rodríguez,. (Fellbaum, 1998b) (Fellbaum, 1998a).

c) DBPedia

DBpedia is a community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to Wikipedia data. It is now almost universally acknowledged that stitching together the world's structured information and knowledge to answer semantically rich queries is one of the key challenges of computer science, and one that is likely to have tremendous impact on the world as a whole. This has led to almost 30 years of research into information integration and ultimately to the Semantic Web and related technologies. Such efforts have generally only gained traction in relatively small and specialized domains, where a closed ontology, vocabulary, or schema could be agreed upon.

However, the broader Semantic Web vision has not yet been realized, and one of the biggest challenges facing such efforts has been how to get enough "interesting" and broadly useful information into the system to make it useful and accessible to a general audience (Auer et al., 2007).

d) Named Entity Recognition

Named Entity Recognition (NER) is to know the entities such as organizations, people, and places from a given text. To ensure that we detect the right entity, we have to know the relationships and the links between entities so we need to know co-reference and anaphoric between them. Terminology extraction is to find the relevant term for a given text such as synonym of an existing work (Elsebai & Meziane, 2011).

1.1. 3 Question Answering

Question Answering (QA) is one of IR field applications. While each engine retrieves documents for a particular query, QA system returns precise answer for a particular query. Most QA systems are evaluated by comparing their results with the results of existing search engines such as Google, Yahoo, and Bing. QA systems work with natural language processing to return information from document by analyzing the natural language of questions to have exact match answer (Kim & Kim, 2008).

1.1. 4 Open Domain Question Answering

In information retrieval, an open domain question answering system aims at returning an answer in response to the user's question. The returned answer is in the form of short texts rather than a list of relevant documents. The system uses three techniques (1) computational linguistics , (2) information retrieval and, (3) knowledge representation for finding answers (Hirschman & Gaizauskas, 2001).

The system takes a natural language question as an input rather than a set of keywords, for example, "When is the national day of Palestine?" This sentence will transform into a query through its logical form. Handling the input in the form of a natural language question makes the system more user-friendly, but it’s hard to implement, as there are various question types and the system will have to identify the correct one in order to give a sensible answer. Assigning a question type to the question is a crucial task, the entire answer extraction process relies on finding the correct question type and hence the correct answer type.

Keyword extraction is an important step for identifying the type of question. In some cases, there are clear words that indicate the question type directly. i.e. "Who", "Where" or "How many", these words tell the system that the answers should be of type "Person", "Location", "Number" respectively. In the example above, the word "When" indicates that the answer should be of type "Date".

Part of Speech tagging and syntactic parsing techniques can also be used to determine the answer type. In this case, the subject is "Palestinian National Day", the predicate is "is" and the adverbial modifier is "when", therefore the answer type is "Date".

Unfortunately, some interrogative words like "Which", "What" or "How" do not give clear answer types (Schmid, 2013).

Each of these words can represent more than one type. In situations like this, other words in the question need to be considered. First thing to do is to find the words that can indicate the meaning of the question. A lexical dictionary such as WordNet (WordNet) can then be used for understanding the context.

For questions such as "Who" or "Where", a Named Entity Recognizer is used to find relevant "Person" and "Location" names from the retrieved documents. Only the relevant paragraphs are selected for ranking.

There are many models that can be used as a strategy for classifying the candidate answers such as Vector Space Model. It Check if the answer is of the correct type as determined in the question type analysis stage. Inference technique can also be used to validate the candidate answers. A weight Number is then given to each of these candidates according to the number of question words it contains and how close these words are to the candidate, the more and the closer the better. The answer is then translated into a compact and meaningful representation by parsing. In the previous example, the expected output answer is "15 Nov 1988"

1.2 Statement of the Problem

Arabic language suffers from the low percentage contents over Wikipedia Articles. Most Arabic language articles are too short to provide encyclopedic coverage of a subject. Such pages are known as „stub" articles in Wikipedia notions (see Figure one for an example). Furthermore, many articles in Arabic Wikipedia lack Infoboxes which are formatted tables usually appear at the top left/right corners of pages. Such an Infobox provides a summary as structured information of some unifying parameters (e.g. name, birth date, age, etc.) about the subject of an article (see Figure 2 for an example).

There are some challenges facing the editing process which affect the average growth of Arabic Wikipedia. Among these challenges are the lack of online textual resources and structured information in Arabic language, the lack of financial resources for editors, and the limitations in time.

Developing method for partially automating the editing process in order to overcome these challenges, would definitely boost the whole process and will help the editors in their task. This method might for example provide resources and some structured information relevant to a particular subject of an article,

In this research, we need to answer these questions: **"how to automatically add Infoboxes and to provide the editors with resources, relevant to particular stub articles in Arabic Wikipedia?"**

1.3 Objectives

The main objective of our research is to develop an automated method that recommends Infobox for Stub Wikipedia Articles in Arabic Language (SWAA). The main objective implies some specific objectives in the following points:

- Data Set: we aim to collect articles from a political domain.
- Implementation: we aim to use kind of techniques and method rooted in Natural language processing, Information retrieval to implement our method for constructing infobox
- Demonstration: we aim to demonstrate the usage on an implemented prototype of our system.
- Evaluation: we aim to evaluate the accuracy of the generated infobox.

The main contribution of this research is a prototype that has the following features:

- it can provide the editors with resources relevant to particular stub Wikipedia Article in Arabic Language.
- it can construct infobox for stub Wikipedia articles in Arabic Language.
- it can generate new Wikipedia articles with basic infobox given to it few key words.

1.4 Signification

We can summarize the importance of the research in the following points:

a) It encourages more research work on topics related to Arabic language retrieval, mining and processing.

b) It encourages editors of Arabic Wikipedia.

c) It contributes in increasing the Arabic language content on the Internet.

d) It can be extended to help in increasing the contents of other Arabic Websites.

1.5 Scope and Limitations

The intended research helps Arab Wikipedia editor to access the relevant articles for the targeted articles to be improved in addition to suggest an Infobox. This affects the content of article and enhance the editing process of Wikipedia.

We can summarize the research scope and limitations in the following points:

a) Our research supports only Arabic language.

b) Our research focuses on constructing Infobox for articles on named entities in specific domain of interest. In this research, we construct infoboxes for named entities in political domain.

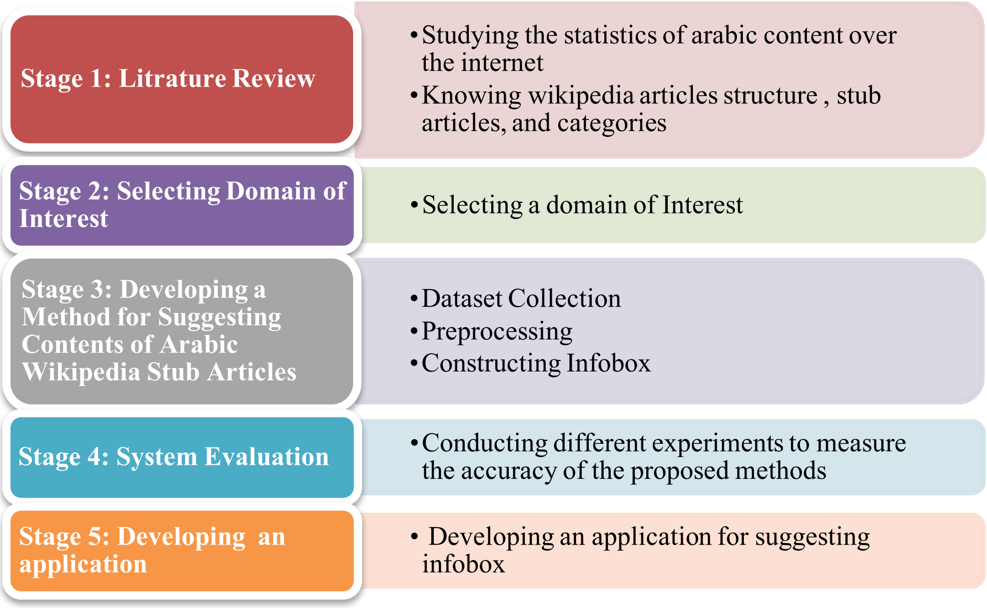
c) The process of collecting data is conducted manually as the development of methods to automatically do that is out of our research focus.

d) The input of our proposed method is only the title or short text if available relevant to a particular stub Wikipedia article.

1.6 Research Methodology

In this research we develop method for suggesting contents of Wikipedia articles either to enrich the contents of existing stub pages or either to generate new ones. The proposed method builds based on existing method in IR, QA, and TM, IE, and NER in order to extract key information from relevant documents on the web. A Wikipedia editing (i.e. mobile application, or portal) that implemented. The automatically generated contents and the different resources from which these contents are extracted suggests for Arabic Wikipedia editors through the proposed editing application/portal. Editors can then make proofreading, revisions, extensions, updates, and/or eliminations of these contents before updating their corresponding articles in Wikipedia. We focus on implementing method for semi-automatic construction of Infoboxes for stub articles in a domain of interest. Infoboxes constructs for articles on named entities in domain of interest.

The research methodology is summarized in **Figure 3.1**. It consists of five main stages, as we shall describe in the following subsections.



Stage 1: Litrature Review

- Studying the statistics of arabic content over the internet
- Knowing wikipedia articles structure , stub articles, and categories

Stage 2: Selecting Domain of Interest

- Selecting a domain of Interest

Stage 3: Developing a Method for Suggesting Contents of Arabic Wikipedia Stub Articles

- Dataset Collection
- Preprocessing
- Constructing Infobox

Stage 4: System Evaluation

- Conducting different experiments to measure the accuracy of the proposed methods

Stage 5: Developing an application

- Developing an application for suggesting infobox

Figure (1.3): Research Methodology

Stage 1: Literature Review

In this stage we investigate the Arabic content over the Internet in order to answer such questions:

“What are the reasons of Arabic content weakness? What is the rate of growth in Arabic content over the internet? What is the rate of growth in other contents compared to Arabic content? “

We answer the above questions by investigating and visualizing many statistical data concerning the Arabic content over the internet particularly in Wikipedia.

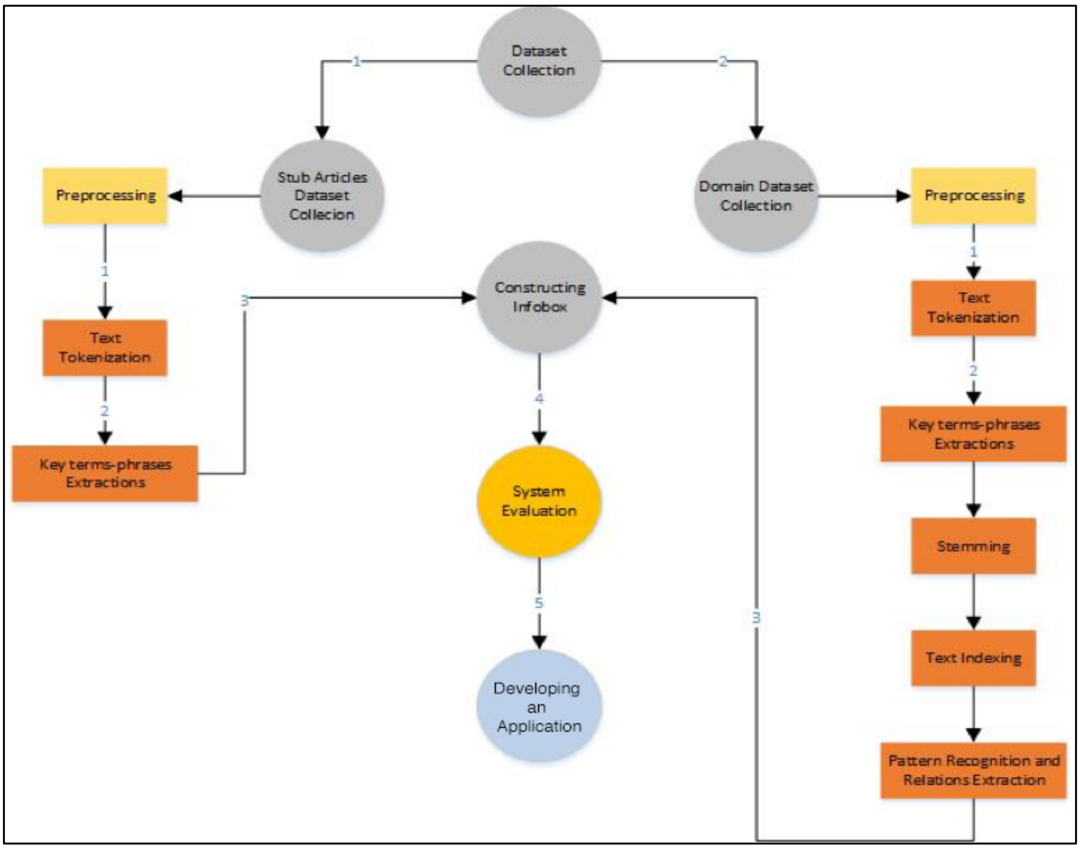
This research applies multiple sources and case studies to know how Wikipedia works and to investigate the structure of the articles. We work with Arabic Wikipedia Dumb and XML parsers to extract the stubs articles.

Stage 2: Selecting Domain of Interest

There are many Arabic articles in Wikipedia in many domains such as politics, society, culture, and sports. In this stage, we select one of these domains to increase and improve its content. The selection process based on the number of Arabic stub articles in each domain so we selected political domain as a case study.

Stage 3: Developing Method for Suggesting Contents for Arabic Wikipedia Stub Articles

We propose **an approach** for constructing Infoboxes for stub articles in a particular domain as the following (see **Figure 4.1** for an overview):



Dataset Collection
1
2
Stub Articles Dataset Collecion
Domain Dataset Collection
Preprocessing
Preprocessing
1
Text Tokenization
2
Key terms-phrases Extractions
3
Constructing Infobox
4
System Evaluation
5
Developing an Application
1
Text Tokenization
2
Key terms-phrases Extractions
Stemming
Text Indexing
Pattern Recognition and Relations Extraction
3

Figure (1.4): The proposed approach for constructing Infoboxes for stub articles

Stage 4: System Evaluation

In order to evaluate the accuracy of the proposed method, we conduct different experiments. A gold standard dataset for Wikipedia articles already having Infoboxes will be created, and then used to measure the accuracy of the Infoboxes generated for the same set of articles by our proposed prototype. Standard measures such accuracy used in order to evaluate the proposed method.

Stage 5: Developing an Application

Finally, we develop an application for extracting information to populate infobox from articles. In this application, editors can do the following:

a) Ask the system to automatically suggested an Infobox for a stub article

b) Update the content of the automatically generated Infobox as needed

c) To facilitate searching and editing process about selected domain by editors.

1.7 Thesis Outlines

The thesis document is structured as follows: In chapter 2 we present the Related Work conducted in the thesis field. We present our proposed approach in chapter 3. The implementation of the applied part of the thesis is presented in chapter 4, while Chapter 5 is about the development and testing expirements and their evaluation. We conclude the thesis in the final chapter 6.

Chapter 2

Related Work

Chapter 2

Related Work

In this chapter we introduce the *Literature Review* related to the scope of our research.

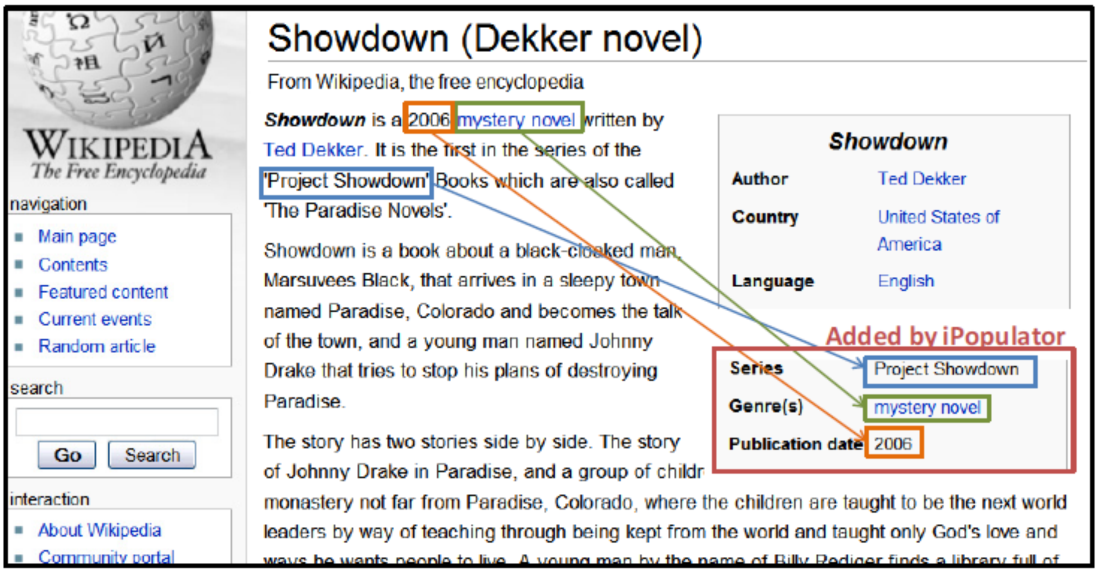
We review many of related works prior works and cite many examples in the fields of Extracting Named Entities, Enriching Wikipedia Content.

2.1 Named Entity Extraction

2.1.1 Extracting Named Entities from Wikipedia Articles

Sriurai,et al.,2009 noted that there are many Wikipedia articles that have outdated info boxes. They conducted a technique to Improve Info boxes of Wikipedia articles by detecting the outdated content of it. They reported that there proposed method extracts new information by combining the pattern-based approach with the entity-search-based approach.

Lange, et al., 2010 found that some info boxes of Wikipedia articles need to be with more content. They introduced a system called iPopulator. It automatically populates the info boxes of Wikipedia articles by extracting attribute values from the article's text. It also detects and exploits the structure of attribute values for independently extracting value parts.



Showdown (Dekker novel)

From Wikipedia, the free encyclopedia

Showdown is a 2006 mystery novel written by Ted Dekker. It is the first in the series of the 'Project Showdown' Books which are also called 'The Paradise Novels'.

Showdown is a book about a black-cloaked man, Marsuvees Black, that arrives in a sleepy town named Paradise, Colorado and becomes the talk of the town, and a young man named Johnny Drake that tries to stop his plans of destroying Paradise.

The story has two stories side by side. The story of Johnny Drake in Paradise, and a group of childr monastery not far from Paradise, Colorado, where the children are taught to be the next world leaders by way of teaching through being kept from the world and taught only God's love and ways he wants people to live. A young man by the name of Billy Rediger finds a library full of

Showdown	
Author	Ted Dekker
Country	United States of America
Language	English

Added by iPopulator

Series	Project Showdown
Genre(s)	mystery novel
Publication date	2006

Figure (2.1): An example for iPopulator

Zhang, et al., 2014 proposed a platform called “WiiCluster”. It’s a scalable platform for automatically generating infoboxes, the system depends on effective cluster algorithm based on a rich of semi-structured Wikipedia articles (linked entities) and it’s effective in generating meaningful summarization of articles and it can generate nearly 10 million facts

The ideas from the mentioned literature is similar to our approach but all of these efforts are focused for English language only. Our intended research concentrates on improving the Arabic language articles.

2.1.2 NER for Arabic Language

Boujelben, et al., 2014 found that there is a high frequency of named entities that do not have any linked information. They also found many studies that are using machine learning approach. They also showed that using a hybrid approach could give more power to the result of extracting relation of Arabic named entity recognition. They obtained promising results by using F-Score measure on their corpus.

Oudah & Shaalan, 2012 concluded that there are many systems were developed using either of the two approaches: rule based model or machine learning with its strength and weakness. They found that the overall performance increased by using a hybrid approach by combining the two approaches mentioned above in a pipeline process.

Wu & Weld, 2008 proposes system called “KOG”, it’s an autonomous system for refining Wikipedia’s infobox-class ontology. This system solves the problem of refinement using machine learning by using both support vector machine and a more powerful joint-inference approach expressed in Markov Logic Networks.

Al Zamil & Al-Radaideh, 2014 proposes a method that extracts ontological relationships. It goals extract semantic features of Arabic text, provide syntactic patterns of relationships among concepts, and a formal model of extracting ontological relations.

Cheddadi, 2014 proposes a method to improve the current question answering performance through surface-based and deeper approaches. The proposed approach divided into three layers: keyword based, structure-based and, semantic-based, this method improves the increase the importance of resource coverage enrichment.

Hammo, et al., 2002 develop a system called (QARAB) that takes Arabic natural language question and provide a short answer, it depends on newspapers articles as a primary source of knowledge from Alraya newspaper in Qatar.

Many researchers made efforts to improve the Wikipedia articles and Arabic Content but not constructing infoboxes, our research is a semi-automatic method for infoboxes extracting from Arabic Wikipedia articles. Thus, we suggest a generated infobox for editor and then he will decide to include it in an article or not.

2.2 Enriching Wikipedia Content

Rothfels, et al., 2011 found that there is a problem of generating recommendations for Wikipedia articles based on constrained data. They also found many approaches for the recommendation systems such as: collaborative filtering techniques, content-based methods, and combine two approaches to predict the user interests. They focused on generating the recommendation system based on content analysis. They reported that their methods are promising for users with many likes but algorithms do not generalize well to more constrained data.

On the other hand, (Sauper & Barzilay, 2009) found that there is a weakness in Wikipedia content. So, they support Wikipedia articles content by creating a multi-paragraph overview article that provides a comprehensive summary of a subject by using text summarization technique which applied on disease synopses articles.

Torres, et al., 2013 studied DBpedia content which is extracted from Wikipedia. They also found that many existing relations among resources in DBpedia are missing links among articles from Wikipedia. They showed that adding these links enrich Wikipedia content by using their algorithm which is called BlueFinder.

Sriurai et al., 2009 found that the current recommendation algorithm of related articles on Wikipedia has a drawback. Current Algorithm missed some links of related articles. They showed that their algorithm could generate a set of recommended articles, which are more relevant than the linked articles given on the test articles.

Yuncong & Fung, 2010 found that the English Wikipedia has many articles but other language of Wikipedia such as Chinese has suffered from the weakness in the content. They showed their approach which is called a synthesis approach. This approach presents the information conveyed by an English article in Chinese language, instead of literally translate it based on a topic-template expressed by the keywords extracted from the English article.

Banerjee & Mitra, 2016 noticed that the increase of Wikipedia content restricted by the availability of authors and editors but, this increase not enough for reader's needs. They introduce a system called (WikiWrite) that generating articles automatically based on machine learning classifier from the similar articles that retrieved from web.

Yuncong & Fung, 2010 noted that current English version of Wikipedia has more than three millions of articles while Chinese version has only one tenth of amount. Chinese articles suffer from content incoherence and lack of details compared to their English counterparts.

They proposed an approach called "synthesis". It is an unsupervised approach for automatically synthesize Wikipedia articles in multiple languages.

In our proposed method, generating infoboxes depends on rule based technique by extracting entities from current Arabic Wikipedia articles and similar articles from the internet. Some of the above approaches use machine learning technique, using machine learning in our approach needs a huge number of Arabic articles as a learning dataset but the current articles on the Internet are not enough for doing that and all of the above researches are focused on English language only. Our intended research concentrates on improving the Arabic language articles.

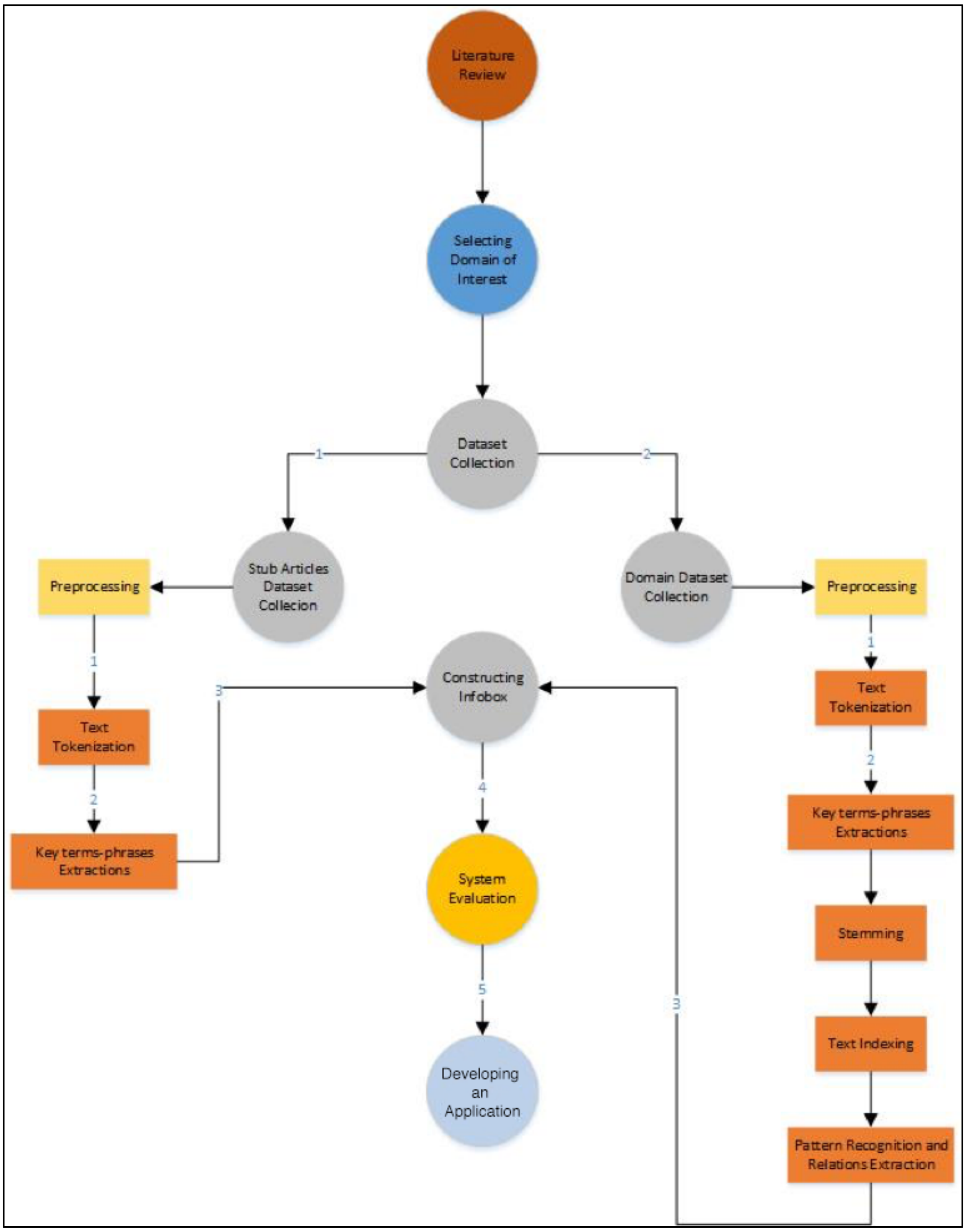
Chapter 3

Proposed Method for Constructing Infobox

Chapter 3

Proposed Methodology for Constructing Infobox

We propose **a method** for constructing Infoboxes for stub articles in a particular domain as we shall describe in the rest of this chapter (see **Figure 3.1** for an overview):



Literature Review
Selecting Domain of Interest
Dataset Collection
1
2
Stub Articles Dataset Collecion
Domain Dataset Collection
Preprocessing
Preprocessing
1
1
Text Tokenization
Text Tokenization
2
2
Key terms-phrases Extractions
Key terms-phrases Extractions
3
Constructing Infobox
4
System Evaluation
5
Developing an Application
Stemming
Text Indexing
Pattern Recognition and Relations Extraction
3

Figure (3.1): The proposed approach for constructing Infoboxes for stub articles

For constructing infobox our method consists of number of phases as follows:

Phase 1: Datasets Collection

In this phase we collect both (1) the stub articles of a particular domain for which we suggest new contents, and (2) the domain dataset from which we extract new contents for the stub articles.

a) Collecting Stub Articles

There are many Arabic articles in Wikipedia in different domains such as politics, society, culture, and sports. In this stage, we collect and process stub articles in Arabic Wikipedia in a given domain. We consider stub articles from Domain of interest.

Wikipedia provides a list of stub articles in any targeted domains.

b) Collecting Domain Dataset

There are many Arabic articles on the Internet relevant to stub articles in Wikipedia. In this sub-phase we search for documents relevant to the domain of interest over the Internet to extract new content for the stub articles.

Phase 2: Preprocessing

Frequently, the texts we collect are not ready for analysis. For example, we need a short text (i.e. query) as an input but the input text so, we have to break up a long text and extract the important key words and terms. Preprocessing is a step that aims at preparing information to be ready for their analysis.

This phase includes the following sub-phases:

a) Text Tokenization

In this sub phase we break up the text (e.g. text of a stub article or text from domain datasets) into tokens/words and symbols.

b) Key Terms/Phrases Extractions

In this sub-phase we identify the key terms/phrases from the text tokens. These key terms/phrases used later in constructing either the search queries or either the search index. Key terms/phrases can include named entities. Therefore, NER techniques uses to extract different named entities from the text. There are three main method of learning NE: Supervised Learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL). The main shortcoming of SL is the requirement of a large annotated corpus. The unavailability of such resources and the prohibitive cost of creating them lead to two other alternative learning methods (Sun, 2010).

1. Supervised Learning

The aim of supervised learning is to study the positive and negative features examples of NE over a large collection of annotated documents and design rules that capture instances of a given type.

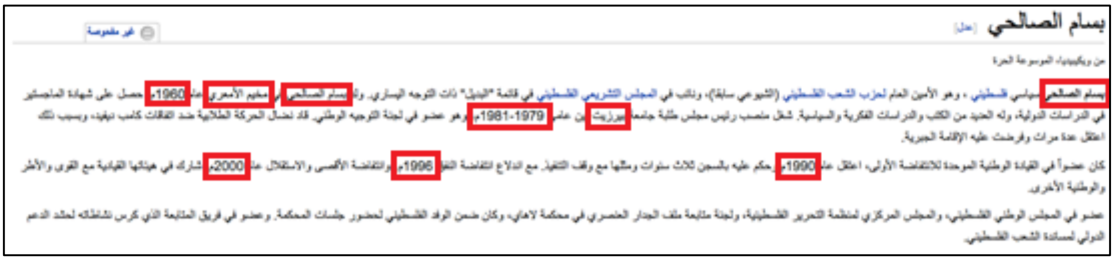
2. Semi-supervised

Semi-supervision is still new. It's depends on a technology that called "bootstrapping" with a small measure of control, such as a row of seeds, for the beginning of the learning process.

3. Unsupervised Learning

Unsupervised learning uses an approach that called “clustering”. For example, one can try to collect names from clustered groups based on the similarity of context and there are other different methods. The techniques depend on lexical resources (e.g. WordNet), calculated on lexical patterns and statistics on a large unannotated corpus.

For example, **Figure 3.2** is for a stub article, and **Figure 3.3** is for a domain of interest article. Named entities that should be extracted from both documents are in red boxes.



بسام الصالحي [عدل] غير مفحوصة

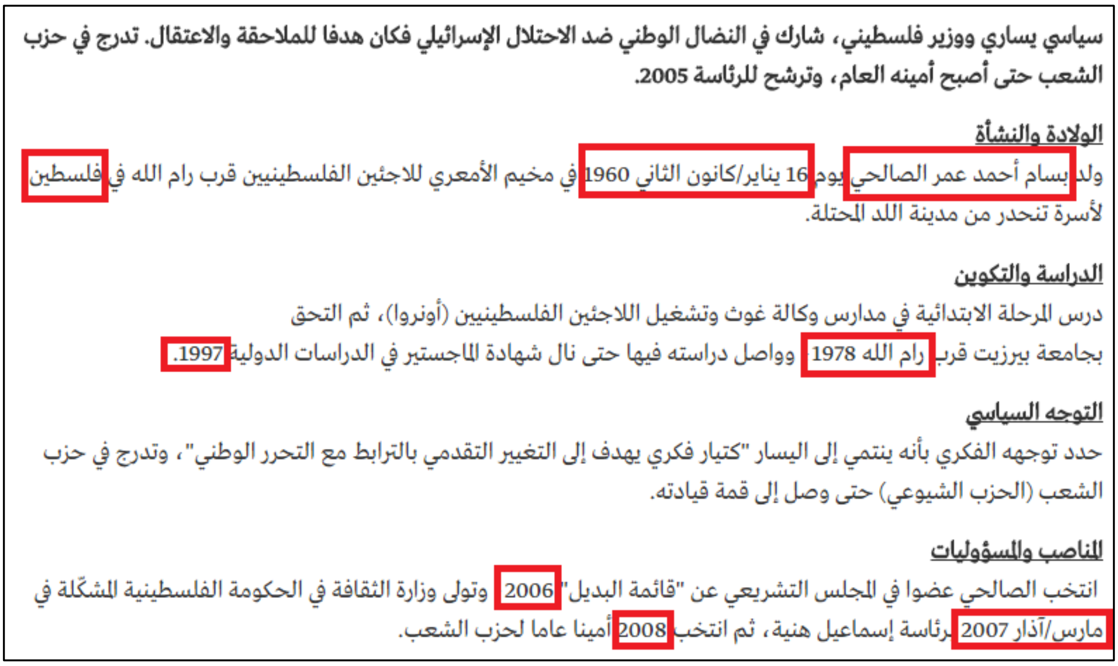
من ويكيبيديا، الموسوعة الحرة

بسام الصالحي سياسي فلسطيني ، وهو الأمين العام لحزب الشعب الفلسطيني (الشيوعي سابقا)، ونائب في المجلس التشريعي الفلسطيني في قائمة "البديل" ذات التوجه اليساري. ولد بسام الصالحي في مخيم الأمعري عام 1960م حصل على شهادة الماجستير في الدراسات الدولية، وله العديد من الكتب والدراسات الفكرية والسياسية. شغل منصب رئيس مجلس طلبة جامعة بيرزيت بين عامي 1979-1981م وهو عضو في لجنة التوجيه الوطني. قاد نضال الحركة الطلابية ضد [illegible]، وبسبب ذلك اعتقل عدة مرات وفرضت عليه الإقامة الجبرية.

كان عضواً في القيادة الوطنية الموحدة للانتفاضة الأولى، اعتقل عام 1990م، حكم عليه بالسجن ثلاث سنوات ومثلها مع وقف التنفيذ. مع اندلاع انتفاضة النفق 1996م وانتفاضة الأقصى والاستقلال عام 2000م، شارك في هيئاتها القيادية مع القوى والأطر الوطنية الأخرى.

عضو في المجلس الوطني الفلسطيني، والمجلس المركزي لمنظمة التحرير الفلسطينية، ولجنة متابعة ملف الجدار العنصري في محكمة لاهاي، وكان ضمن الوفد الفلسطيني لحضور جلسات المحكمة. وعضو في فريق المتابعة الذي كرس نشاطه لحشد الدعم الدولي لمساندة الشعب الفلسطيني.

Figure (3.2): Named entities extracted from a stub article



سياسي يساري ووزير فلسطيني، شارك في النضال الوطني ضد الاحتلال الإسرائيلي فكان هدفا للملاحقة والاعتقال. تدرج في حزب الشعب حتى أصبح أمينه العام، وترشح للرئاسة 2005.

الولادة والنشأة

ولد بسام أحمد عمر الصالحي يوم 16 يناير/كانون الثاني 1960 في مخيم الأمعري للاجئين الفلسطينيين قرب رام الله في فلسطين لأسرة تنحدر من مدينة اللد المحتلة.

الدراسة والتكوين

درس المرحلة الابتدائية في مدارس وكالة غوث وتشغيل اللاجئين الفلسطينيين (أونروا)، ثم التحق بجامعة بيرزيت قرب رام الله 1978 وواصل دراسته فيها حتى نال شهادة الماجستير في الدراسات الدولية 1997.

التوجه السياسي

حدد توجهه الفكري بأنه ينتمي إلى اليسار "كتيار فكري يهدف إلى التغيير التقدمي بالترابط مع التحرر الوطني"، وتدرج في حزب الشعب (الحزب الشيوعي) حتى وصل إلى قمة قيادته.

المناصب والمسؤوليات

انتخب الصالحي عضوا في المجلس التشريعي عن "قائمة البديل" 2006 وتولى وزارة الثقافة في الحكومة الفلسطينية المشكّلة في مارس/آذار 2007 برئاسة إسماعيل هنية، ثم انتخب 2008 أمينا عاما لحزب الشعب.

Figure (3.3): Named entities extracted from a domain of interest article

(c) Stemming

Stemming phase is used to extract the sub-part i.e. called as stem/root of a given word. For example, the words “ولد", "ولادة", "ميلاد" all can be rooted to the word "ولد". The main role of stemming is to remove various suffixes as result in the reduction of number of words, to have exactly matching stems,

- **ولد** سليم في فلسطين
- وكانت **ولادة** سليم في فلسطين
- سليم **مولود** في فلسطين

On the completion of stemming process, next step is to count the frequency of each word. Information retrieval works on the output of this tokenization process for achieving or producing most relevant results to the given "ولد"

(d) Text Indexing

Indexing is a technique aims at finding the results of search query easily. It is similar to an index at the back of a book. In our research, we have a relevant articles (Documents) in our dataset. Indexing helps in look up for search term and find corresponding articles in documents. When we create index based document, we can find out the matched document of search query. Indexing technique used by search engines to facilitate fast and accurate information retrieval.

The following steps explain how indexing process:

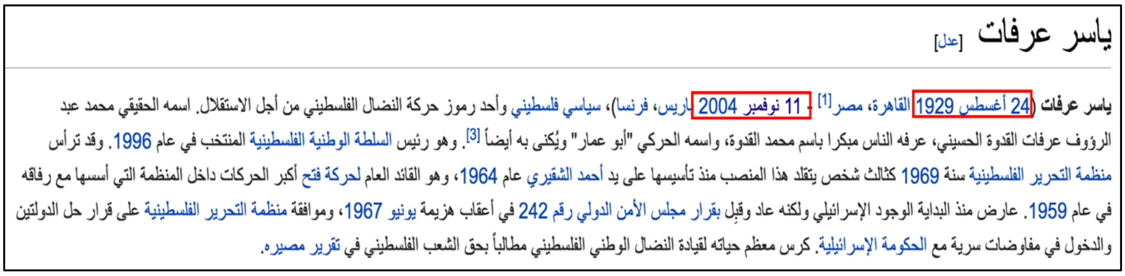
1. Preparing text files.
2. Giving each text file an index.
3. Adding text files as documents.
4. Adding each documents to its index.
5. Giving the engine a search query.
6. Search engine searching for the index and return the matched document that have a search query.

All of the above steps can be done by tools such as Lucene.

(e) Collecting Relations

Our goal of this step is to detect named entities that related to Full Name, (Birth, Death) Locations, and (Birth, Death) Dates from text.

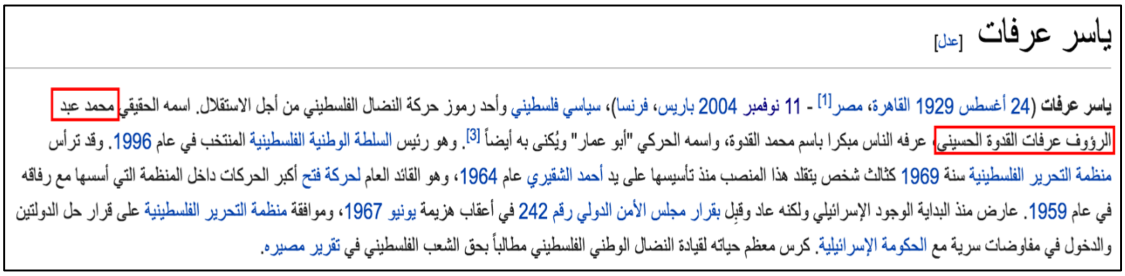
To achieve this goal, we collect a common pattern that appeared with each named entity, all of these patterns are collected manually based on exploring different articles from many resources as **Figures 3.4, Figure 3.5 and Figure 3.6**:



ياسر عرفات [عدل]

ياسر عرفات (24 أغسطس 1929 القاهرة، مصر [1] - 11 نوفمبر 2004 باريس، فرنسا)، سياسي فلسطيني وأحد رموز حركة النضال الفلسطيني من أجل الاستقلال. اسمه الحقيقي محمد عبد
الرؤوف عرفات القدوة الحسيني، عرفه الناس مبكرا باسم محمد القدوة، واسمه الحركي "أبو عمار" ويُكنى به أيضاً. [3] وهو رئيس السلطة الوطنية الفلسطينية المنتخب في عام 1996. وقد ترأس
منظمة التحرير الفلسطينية سنة 1969 كثالث شخص يتقلد هذا المنصب منذ تأسيسها على يد أحمد الشقيري عام 1964، وهو القائد العام لحركة فتح أكبر الحركات داخل المنظمة التي أسسها مع رفاقه
في عام 1959. عارض منذ البداية الوجود الإسرائيلي ولكنه عاد وقبِل بقرار مجلس الأمن الدولي رقم 242 في أعقاب هزيمة يونيو 1967، وموافقة منظمة التحرير الفلسطينية على قرار حل الدولتين
والدخول في مفاوضات سرية مع الحكومة الإسرائيلية. كرس معظم حياته لقيادة النضال الوطني الفلسطيني مطالباً بحق الشعب الفلسطيني في تقرير مصيره.

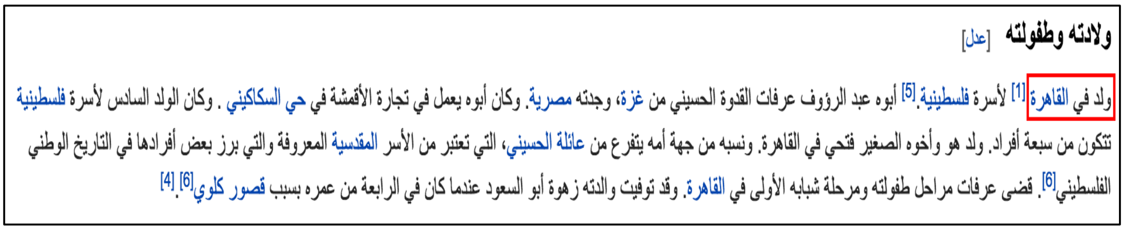
Figure (3.4): Birth and Death Date Pattern Sample (Wikipedia).



ياسر عرفات [عدل]

ياسر عرفات (24 أغسطس 1929 القاهرة، مصر [1] - 11 نوفمبر 2004 باريس، فرنسا)، سياسي فلسطيني وأحد رموز حركة النضال الفلسطيني من أجل الاستقلال. اسمه الحقيقي محمد عبد
الرؤوف عرفات القدوة الحسيني، عرفه الناس مبكرا باسم محمد القدوة، واسمه الحركي "أبو عمار" ويُكنى به أيضاً. [3] وهو رئيس السلطة الوطنية الفلسطينية المنتخب في عام 1996. وقد ترأس
منظمة التحرير الفلسطينية سنة 1969 كثالث شخص يتقلد هذا المنصب منذ تأسيسها على يد أحمد الشقيري عام 1964، وهو القائد العام لحركة فتح أكبر الحركات داخل المنظمة التي أسسها مع رفاقه
في عام 1959. عارض منذ البداية الوجود الإسرائيلي ولكنه عاد وقبِل بقرار مجلس الأمن الدولي رقم 242 في أعقاب هزيمة يونيو 1967، وموافقة منظمة التحرير الفلسطينية على قرار حل الدولتين
والدخول في مفاوضات سرية مع الحكومة الإسرائيلية. كرس معظم حياته لقيادة النضال الوطني الفلسطيني مطالباً بحق الشعب الفلسطيني في تقرير مصيره.

Figure (3.5): Full Named Pattern Sample (Wikipedia).



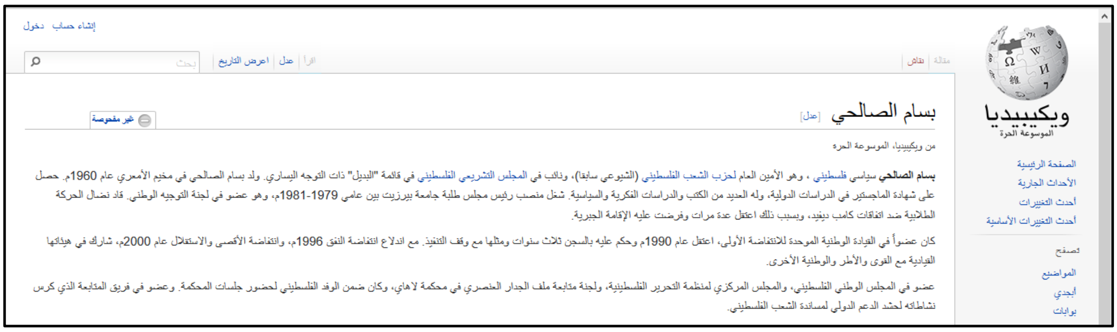
ولادته وطفولته [عدل]

ولد في القاهرة [1] لأسرة فلسطينية. [5] أبوه عبد الرؤوف عرفات القدوة الحسيني من غزة، وجدته مصرية. وكان أبوه يعمل في تجارة الأقمشة في حي السكاكيني . وكان الولد السادس لأسرة فلسطينية
تتكون من سبعة أفراد. ولد هو وأخوه الصغير فتحي في القاهرة. ونسبه من جهة أمه يتفرع من عائلة الحسيني، التي تعتبر من الأسر المقدسية المعروفة والتي برز بعض أفرادها في التاريخ الوطني
الفلسطيني[6]. قضى عرفات مراحل طفولته ومرحلة شبابه الأولى في القاهرة. وقد توفيت والدته زهوة أبو السعود عندما كان في الرابعة من عمره بسبب قصور كلوي[6].[4]

Figure (3.6): Brith Location Pattern Sample (Wikipedia).

Phase 3: Constructing Infobox

In this phase, we aim at constructing infobox for a stub article. Suppose that we want to enhance the stub article shown in **Figure 3.7:**



بسام الصالحي [عدل]

من ويكيبيديا، الموسوعة الحره

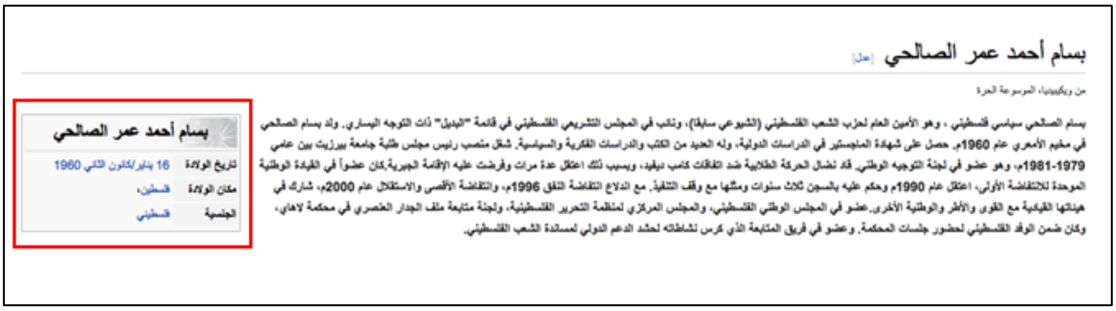
بسام الصالحي سياسي فلسطيني ، وهو الأمين العام لحزب الشعب الفلسطيني (الشيوعي سابقا)، ونائب في المجلس التشريعي الفلسطيني في قائمة "البديل" ذات التوجه اليساري. ولد بسام الصالحي في مخيم الأمعري عام 1960م. حصل
على شهادة الماجستير في الدراسات الدولية، وله العديد من الكتب والدراسات الفكرية والسياسية. شغل منصب رئيس مجلس طلبة جامعة بيرزيت بين عامي 1979-1981م، وهو عضو في لجنة التوجيه الوطني. قاد نضال الحركة
الطلابية ضد اتفاقات كامب ديفيد، وبسبب ذلك اعتقل عدة مرات وفرضت عليه الإقامة الجبرية.

كان عضواً في القيادة الوطنية الموحدة للانتفاضة الأولى، اعتقل عام 1990م وحكم عليه بالسجن ثلاث سنوات ومثلها مع وقف التنفيذ. مع اندلاع انتفاضة النفق 1996م، وانتفاضة الأقصى والاستقلال عام 2000م، شارك في هيئاتها
القيادية مع القوى والأطر والوطنية الأخرى.

عضو في المجلس الوطني الفلسطيني، والمجلس المركزي لمنظمة التحرير الفلسطينية، ولجنة متابعة ملف الجدار العنصري في محكمة لاهاي، وكان ضمن الوفد الفلسطيني لحضور جلسات المحكمة. وعضو في فريق المتابعة الذي كرس
نشاطاته لحشد الدعم الدولي لمساندة الشعب الفلسطيني.

Figure (3.7): Example of Stub Article

We expect to construct an info box that contains named entities as the shown in **Figure 3.8:**



بسام أحمد عمر الصالحي [عدل]

من ويكيبيديا، الموسوعة الحرة

بسام الصالحي سياسي فلسطيني ، وهو الأمين العام لحزب الشعب الفلسطيني (الشيوعي سابقا)، ونائب في المجلس التشريعي الفلسطيني في قائمة "البديل" ذات التوجه اليساري. ولد بسام الصالحي في مخيم الأمعري عام 1960م. حصل على شهادة الماجستير في الدراسات الدولية، وله العديد من الكتب والدراسات الفكرية والسياسية. شغل منصب رئيس مجلس طلبة جامعة بيرزيت بين عامي 1979-1981م، وهو عضو في لجنة التوجيه الوطني. قاد نضال الحركة الطلابية ضد اتفاقات كامب ديفيد، وبسبب ذلك اعتقل عدة مرات وفرضت عليه الإقامة الجبرية.كان عضواً في القيادة الوطنية الموحدة للانتفاضة الأولى، اعتقل عام 1990م وحكم عليه بالسجن ثلاث سنوات ومثلها مع وقف التنفيذ. مع اندلاع انتفاضة النفق 1996م، وانتفاضة الأقصى والاستقلال عام 2000م، شارك في هيئاتها القيادية مع القوى والأطر والوطنية الأخرى.عضو في المجلس الوطني الفلسطيني، والمجلس المركزي لمنظمة التحرير الفلسطينية، ولجنة متابعة ملف الجدار العنصري في محكمة لاهاي، وكان ضمن الوفد الفلسطيني لحضور جلسات المحكمة. وعضو في فريق المتابعة الذي كرس نشاطاته لحشد الدعم الدولي لمساندة الشعب الفلسطيني.

بسام أحمد عمر الصالحي	
تاريخ الولادة	16 يناير/كانون الثاني 1960
مكان الولادة	فلسطين،
الجنسية	فلسطيني

Figure (3.8): Example of Generated Infobox

Below is a scenario that explain the steps of constructing info box:

a. Editor will enter the search query to start searching about relevant articles. Suppose that the short input was “بسام الصالحي ولادة و نشأة “.

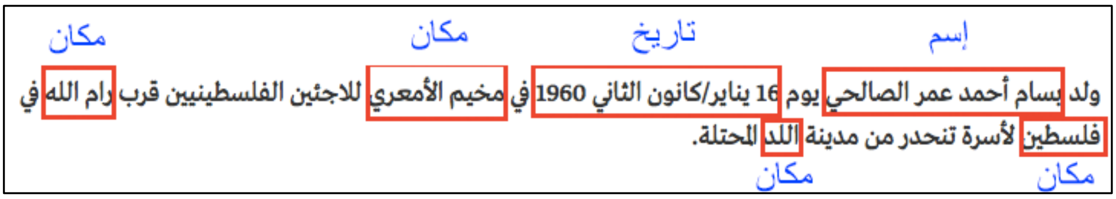
b. During the preprocessing phase of stub articles, we have **a tokenization key/term phrases and NER** steps. We need these steps to get only the name from a search query.

c. The search query will be "بسام الصالحي".

d. Our system will search for input query in the stub articles dataset, especially in named entities (Persons).

e. This query will be used to search for relevant documents using a search engine such as Lucene**.**

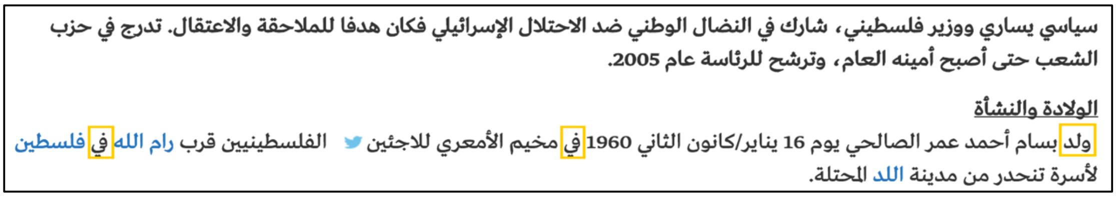
f. Our system will extract named entities from the returned document and then detecting Patterns as **Figure 3.9**:



ولد بسام أحمد عمر الصالحي يوم 16 يناير/كانون الثاني 1960 في مخيم الأمعري للاجئين الفلسطينيين قرب رام الله في فلسطين لأسرة تنحدر من مدينة اللد المحتلة.

Figure (3.9): Example of Named Entity Recognition

g. Extracting relation from the detected named entities to get the right and targeted named entities as **Figure 3.10**, **Figure 3.11**:

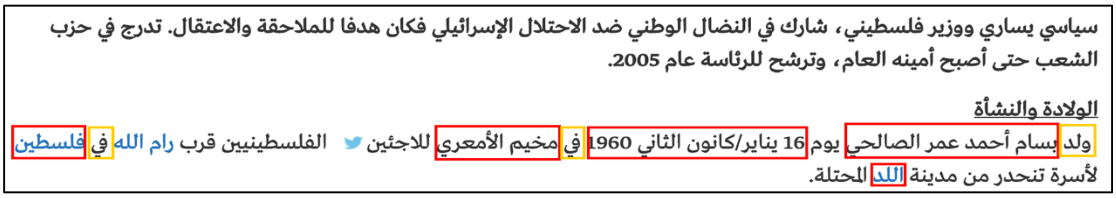


سياسي يساري ووزير فلسطيني، شارك في النضال الوطني ضد الاحتلال الإسرائيلي فكان هدفا للملاحقة والاعتقال. تدرج في حزب الشعب حتى أصبح أمينه العام، وترشح للرئاسة عام 2005.

الولادة والنشأة

ولد بسام أحمد عمر الصالحي يوم 16 يناير/كانون الثاني 1960 في مخيم الأمعري للاجئين الفلسطينيين قرب رام الله في فلسطين لأسرة تنحدر من مدينة اللد المحتلة.

Figure (3.10): Example of Detecting Relations



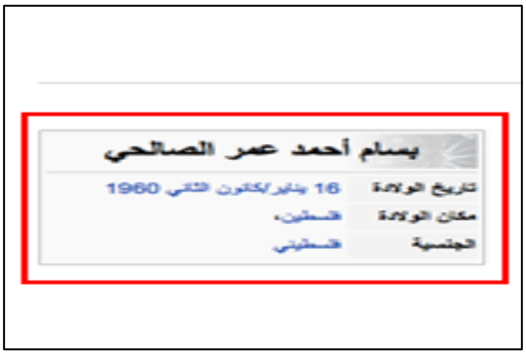
سياسي يساري ووزير فلسطيني، شارك في النضال الوطني ضد الاحتلال الإسرائيلي فكان هدفا للملاحقة والاعتقال. تدرج في حزب الشعب حتى أصبح أمينه العام، وترشح للرئاسة عام 2005.

الولادة والنشأة

ولد بسام أحمد عمر الصالحي يوم 16 يناير/كانون الثاني 1960 في مخيم الأمعري للاجئين الفلسطينيين قرب رام الله في فلسطين لأسرة تنحدر من مدينة اللد المحتلة.

Figure (3.11): Example of Detecting Related Entities

h. The named entities will be ready to construct infobox as **Figure 3.12**:



بسام أحمد عمر الصالحي	
تاريخ الولادة	16 يناير/كانون الثاني 1960
مكان الولادة	فلسطين،
الجنسية	فلسطيني

Figure (3.12): Example of Expected Infobox

In order to fill the gabs in the infobox, we designed a set of algorithms, we divide our entities recognition algorithm into the sub algorithms as a following:

A. Full Name

In some articles of Wikipedia, the name of article doesn't reflect the real name of entity for example, Yasser Arafat is the name of article but his real name is (محمد عبد الرؤوف القدوة), in this step we detect the full name of targeted entity from article as the following:

1. Extracting named entities (Person) from text.
2. Segment text to words.
3. Give each entity index.
4. Search for entities that have a difference of one with the previous entity.

5. Add entities to a list.
6. Show full name.

Table (3.1): Full Name Detection Algorithm

```
input: paragraph segmented to words, persons (entities)


output: Full name candidate
index = 0
for each entity in entities List:
    give entity an index
    index++
get entities based on indexes
sort entities ascending
```

Below is an example that explain the steps of algorithm:

" لا يعرف على وجه اليقين مكان ولادة محمد عبد الرؤوف القدوة الحسيني الذي اشتهر فيما بعد باسم ياسر عرفات أو أبو عمار "

1- Segment the paragraph

لا، يعرف ، على ، وجه ، اليقين ، مكان ، ولادة ، محمد ، عبد ، الرؤوف ، القدوة ، الحسيني ، الذي ، اشتهر ، فيما ، بعد ، باسم ، ياسر ، عرفات ، أبو ، عمار

2- Extract Named Entities (Person)

محمد ، عبد ، الرؤوف ، القدوة ، الحسيني ، ياسر ، عرفات ، أبو عمار

3- Give each word an index

لا 0 ، يعرف 1، على 2، وجه 3، اليقين 4، مكان 5، ولادة 6، محمد 7، عبد 8، الرؤوف 9، القدوة 10،

الحسيني 11، الذي 12، اشتهر 13، فيما 14، بعد 15، باسم 16، ياسر 17، عرفات 18، أبو 19، عمار 20

4- Calculate distance between entities

لا 0، يعرف 1، على 2، وجه 3، اليقين 4، مكان 5، ولادة 6، محمد 7، عبد 8، الرؤوف 9، القدوة 10،

الحسيني 11، الذي 12، اشتهر 13، فيما 14، بعد 15، باسم 16، ياسر 17، عرفات 18، أبو 19، عمار 20

The distance between each entity will be 1 as below:

Table (3.2): Calculate distance between entities

Next Word	Previous Word	Distance	Status
عبد	محمد	abs (7-8) = 1	T
الرؤوف	عبد	abs (8-9) = 1	T
القدوة	الرؤوف	abs (9-10) = 1	T
الحسيني	القدوة	abs (10-11) = 1	T
الحسيني	ياسر	abs (17-11) = 6	F

5 – Show the Final Result

محمد عبد الرؤوف القدوة الحسيني

B. Location (Birth Location, Death Location)

In this step we detect the entities location that reflect the birth and death location of person as the following:

1. Extracting named entities (GPE) from text.
2. Segment text to words.
3. Give each word and index.
4. Searching for sentence that contain keywords (نشأ ، ولد ، مات ، توفي)
5. Calculate distance between keywords and (GPE) entities
6. Show entity with minimum distance as a candidate.

Table (3.3): Location Detection Algorithm

```
input: Extracted entities (GPE), Paragraph words, and, keyword
(نشأ ، ولد ، مات ، توفي)
output: Birth Location, Death Location
min = 0
for each word in paragraph sentences:
    Search for keyword (وُلَد) in paragraph:
        If search result == 0:
            min = -1
            if extracted entities == 1:
                Show first entity as a candidate
            else if extracted entities > 1:
                Give each entity index e.g. (0) for first entity
                Show the entity that has min index as a candidate
        Else If search result > 0:
            Call simpleRanking()
            Show best candidate based on minimum distance
```

Below is an example that explain the steps of algorithm:

" لا يعرف على وجه اليقين مكان ولادة محمد عبد الرؤوف القدوة الحسيني الذي اشتهر فيما بعد باسم ياسر عرفات أو أبو عمار، والأغلب أنه ولد في القاهرة في الرابع والعشرين من أغسطس/ آب عام 1929، وهو الابن السادس لأب كان يعمل في التجارة، وهاجر إلى القاهرة عام 1927 وعاش في حي السكاكيني. وعندما توفيت والدته وهو في الرابعة من عمره أرسله والده إلى القدس، وهناك بدأ وعيه يتفتح على أحداث ثورة 1936."

1. Search for " ولد ، نشأ ، مات ، توفي " using Lucene
2. The result of search for keywords is as below:

والأغلب أنه ولد في القاهرة في الرابع والعشرين من أغسطس/ آب عام 1929 وهو الابن السادس لأب كان يعمل في التجارة، وهاجر إلى القاهرة عام 1927 وعاش في حي السكاكيني.

3. Detect named entities (Location)

القاهرة

4. Segment the Paragraph

و ، الاغلب ، ان ، ه ، ولد ، في ، القاهرة ، في ، الرابع ، و ، العشرين ، من ، أغسطس ، اب ، عام ،
، و ، هو ، الابن ، السادس ، لاب ، كان ، يعمل ، في ، التجارة ، و ، هاجر ، الى ، القاهرة ، عام
،1927 ، عاش ، في ،حي ، السكاكيني

5. Give each word an index

و 0 ، الاغلب 1 ، ان 2 ، ه 3 ، ولد 4 ، في 5 ، القاهرة 6 ، في 7 ، الرابع 8 ، و 9 ، العشرين 10 ، من
11، أغسطس 12 ، اب 13 ، عام 14، 1929 15، و 16 ، هو 17 ، الابن 18 ، السادس 19 ، لاب 20 ،
كان 22، يعمل 23 ، في 24 ، التجارة 25 ، و 26 ، هاجر 27 ، الى 28 ، القاهرة 29 ، عام 30 ،1927
31، عاش 32 في 33 ،حي 34 ، السكاكيني 35 .

6. Calculate Minimum distance between key word (ولد) and Location Entity

Table (3.4): Calculate Minimum distance between key word (ولد) and Location Entity

Key Word	Location Entity	Distance	Status
ولد	القاهرة	abs (4-6) = 2	T

C. Date (Birth Date, Death Date)

In this section we detect the birth date and the death date by matching date patterns based on regular expression of date as the following:

1. Define date expressions.
2. Search for expressions in the text.
3. Show first date as a birth date.
4. Show second date as a death date

Table (3.5): Date Detection Algorithm

```
input: paragraph, date expressions
output: Birth Date, Death Date
For each sentence in paragraph
        If date exist and match date expressions:
                Then add to list
If result == 1:
        Show date as a birth date
Else if result > 1:
        Show first date as a birth date
        Show second date as a death date
```

Below is an example that explain the steps of algorithm:

ياسر عرفات (24أغسطس 1929 القاهرة، مصر[1] 11نوفمبر 2004 باريس، فرنسا)، سياسي فلسطيني وأحد
رموز حركة النضال الفلسطيني من أجل الاستقلال. اسمه الحقيقي محمد عبد الرؤوف عرفات القدوة الحسيني،
عرفه الناس مبكرا باسم محمد القدوة، واسمه الحركي "أبو عمار" ويُكنى به أيضاً.

1- Lucene searches for a regex pattern as below:
    a. /(\d{1,2})(\s)(.{4,15})(\s)(\d{2,4})/g
2- Lucene shows the result of death and birth date

ياسر عرفات (24أغسطس 1929 القاهرة، مصر[1] 11 -نوفمبر 2004 باريس، فرنسا

D. Ranking

A ranking is a relationship between a set of items such that, for any two items, the first is either 'ranked higher than', 'ranked lower than' or 'ranked equal to' the second. In our proposed method we developed two ranking algorithms, the first is simple ranking that aims at calculating the minimum distance between keywords such as (ولد،نشأ،مات،توفي) in birth, death location entity detection.

Simple Ranking

We use this algorithm inside location algorithm to show the best candidate by calculating a minimum distance between keyword and location entities as follows:

1.1 Segment text to words.

1.2 Giving index for each word in the sentence.

1.3 Calcuate distance between keyword and entities.

1.4 Show the entity that has minimum distance value

Table (3.6): Ranking Algorithm

```
input: Word No which has keyword e.g. (1), extracted entities and, paragraph sentences
output: the distance between keyword and location entity
wordNo = min = 1
for each entity in sentences:
        Give each entity index e.g. (0) for first entity
Calculate distance between word and all entities
Show the minimum value as a candidate
```

Chapter 4

System Technical Implementation

Chapter 4

System Technical Implementation

This chapter explores the technical side of how we accomplished system and gives preface about the tools and packages we used. Each part of our prototype is programmed and implemented using java programming language. Minute details about how we implement specific part in the following subsections.

4.1 Hardware and software specifications

In the following subsection a brief detail about the hardware specification we used in our experiments, in addition to the software tools and packages used while implementing our sentiment analysis system.

4.1.1 Hardware specifications

The machine specifications we used is a MacBook Pro Laptop with 2.5 GHz Intel Core i7 processor, 4 cores, with 16 GB physical memory. Supported with hard disk with 512 GB.

4.1.2 Software Specifications

4.1.2.1 Java and Netbeans IDE

Netbeans is an integrated development environment (IDE), used to develop applications in many other programming languages not just Java. Java is an object-oriented functional computer programming language that enable programmers to develop their applications and not giving concerns on which platform the application will run. As it complied with bytecode that can run on any Java virtual machine (JVM) regardless of computer architecture (Netbeans).

4.1.2.2 Lucene

It is an open source java software library for indexing and searching, it stores each piece of data as document is essentially a collection of fields. LUCENE provides a dynamic index and supports with highly expressive search API to index and retrieving documents from the index (Apache).

4.1.2.3 Arabic Toolkit Service

It is a set of NLP components targeting Arabic language. It’s developed by Microsoft Advanced Technology Lab in Cairo, The component suite includes a full-fledged morphological analyzer (SARF), a spell-checker, an auto corrector, a diacritizer, a named entity recognizer (NER), a colloquial to Arabic converter, and a part-of-speech (POS) tagger.

These components are integrated into multiple Microsoft products and services, such as Windows, Office, Bing, Exchange, SharePoint, and Windows Phone. The ATKS avails these components in the form of web services and associated APIs hosted on Windows Azure (ATKS).

4.1.2.4 GATE

GATE is an open source software capable of solving almost any text processing problem, a mature and extensive community of developers, users, educators, students and scientists , a defined and repeatable process for creating robust and maintainable text processing workflows , in active use for all sorts of language processing tasks and applications, including: voice of the customer; cancer research; drug research ,decision support ,recruitment ,web mining , information extraction and semantic annotation (GATE).

4.1.2.5 Stanford Arabic Word Segmenter & Arabic Tokenizer

Tokenization of raw text is a standard pre-processing step for many NLP tasks. Arabic is a root and template language with abundant bound clitics. These clitics include possessives, pronouns, and discourse connectives. The Stanford Word Segmenter currently supports Arabic. The provided segmentation schemes have been found to work well for a variety of applications.

A tokenizer divides text into a sequence of tokens, which roughly correspond to "words". They provide a class suitable for Arabic tokenization called ArabicTokenizer (Stanford).

4.1.2.6 LingPipe

LingPipe is tool kit for processing text using computational linguistics. It is used to do tasks like Find the names of people, organizations or locations in news, automatically classify Twitter search results into categories and Suggest correct spellings of queries (LingPipe).

4.1.2.7 JWPL

Java Wikipedia Library is a Java-based application programming interface that allows to access all information in Wikipedia. It is fast and efficient access to Wikipedia, Parser for the MediaWiki syntax and, Language independent, in addition to the core functionality, JWPL allows access to Wikipedia's edit history with the Wikipedia Revision Toolkit (JWPL).

4.2 Framework Implementation

As we gave details about our approach in chapter 3 and details about hardware and software specification used in implementing our proposed system in section 4.1. We want to put all together. All software development done by using Netbeans which is an integrated development environment (IDE). Netbeans includes full support for the java Platform Standard Edition Version 8, and also supported Java run time environment JRE8.

4.2.1 Preprocessing

In this subtask different packages were used, Lucene, ATKS, GATE and Arabic Tokenizer. There are five main implementations for this subtask:

1. **Tokenize text:** The first step in preprocessing is to tokenize text. this is accomplished by the Stanford Arabic tokenization it contains *ArabicTokenizer* and *TokenizerFactory* classes, *TokenizerFactory* takes text as an input a return a text a list of strings after tokenization.
2. **Key Terms / Phrases Extraction:** ATKS and GATE are used to extract named entities from text.
3. **Stemming:** Arabic light stemmer from apache Lucene used for getting the root/stem for given words in the text.
4. **Text Indexing:** Lucene used in this task for indexing the dataset to speed up searching into text. This is done using set of classes such as *IndexWriter*, *IndexReader*, *Analyzer*, *Document*, and *TopScoreDocCollector*. 1) Search the

indexed dataset for given word using *IndexSearcher* class and *QueryParser* to compose the search query. 2) The search result will be returned as a list of documents.

5. **Collecting Relations:** Arabic work segment used in this task to segment the result of first document that returned from Lucene and then our system calculates the minimum distance between given word such as “ولد” and the extracted named entities. This is done using *ArabicSegmenter* class.

4.2.2 Constructing Infobox

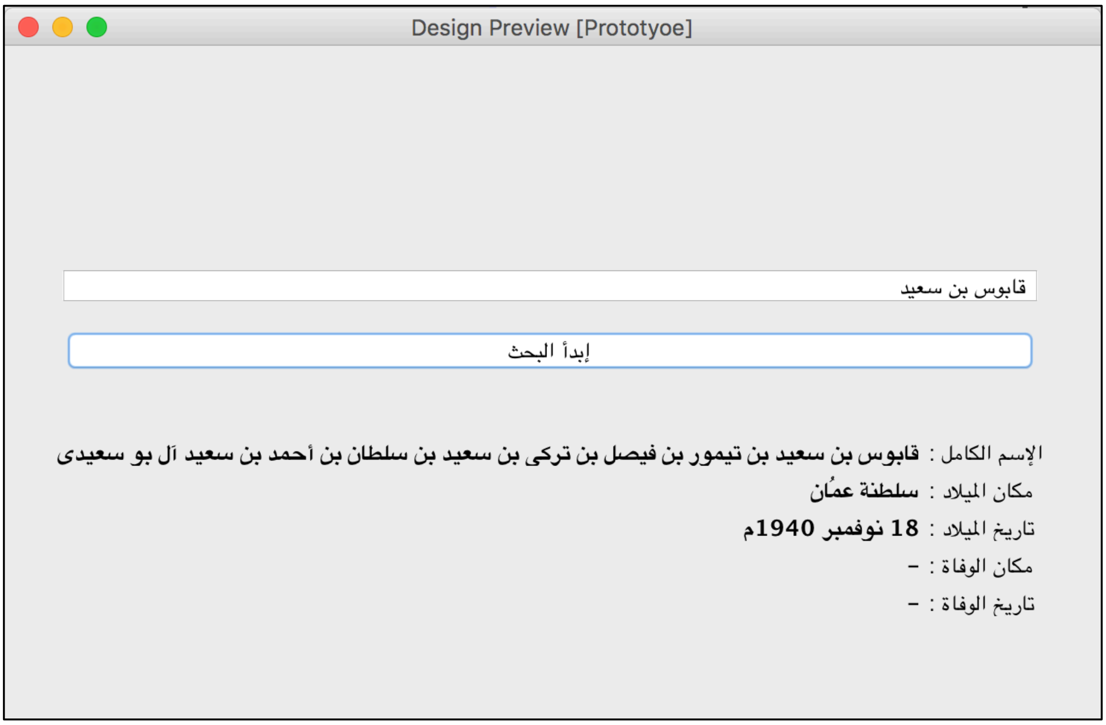
In this subtask different packages were used, ATKS, GATE and LingPipe. There are three main implementations for this subtask:

1. **Extracting Full Name:** Gate and ATKS used in this task to detect full name but we prefer use ATKS because it has a large dataset and gives more accurate results than Gate. This is done using set of classes such as *ArrayOfNamedEntity*, and *NamedEntity*.
2. **Extracting Birth and Death Location:** Gate and ATKS used in this task to detect birth and death location by calculating a minimum distance between keyword and location entities but we prefer use ATKS because it has a large dataset and gives more accurate results than Gate.
3. **Extracting Birth and Death Date:** LingPipe used in this task to detect dates patterns and this done by using a set of classes such as *RegExChunker*, and *Chunking.*

4.2.3 System Demonstration

Below is a brief demonstration of how our system prototype works:

1. Editor will enter the search query to start searching about relevant articles. Suppose that the short input was “قابوس بن سعيد ولادة و نشأة “.
2. During the preprocessing phase of stub articles, we have **a tokenization key/term phrases and NER** steps. We need these steps to get only the name from a search query.
3. The search query will be "قابوس بن سعيد ".



قابوس بن سعيد

الإسم الكامل : قابوس بن سعيد بن تيمور بن فيصل بن تركى بن سعيد بن سلطان بن أحمد بن سعيد آل بو سعيدى

مكان الميلاد : سلطنة عمُان

تاريخ الميلاد : 18 نوفمبر 1940م

مكان الوفاة : –

تاريخ الوفاة : –

Figure (4.1): Screen shot of our prototype

4. Our system will search for input query in the stub articles dataset, especially in named entities (Persons).
5. This query will be used to search for relevant documents from a search engine such as **Lucene.**
6. Our system will extract named entities from the returned document and then detecting Patterns.
7. Extracting relation from the detected named entities to get the right and targeted named entities.
8. Our system will construct infobox and suggest it to editor.

Chapter 5

Results and Discussion

Chapter 5

Results and Discussion

We want to evaluate our implemented approach for constructing Infoboxes for Wikipedia stub articles. We evaluate different components of our algorithm mainly the components presented in Section 5.1, 5.2, 5.3, 5.4 and 5.5. In the following sections, we describe each component dataset, input/output, experiment settings, result, and evaluation metrics.

5.1 Datasets

In order to evaluate our prototype, we constructed two gold standard datasets. The two datasets are collections of articles in Arabic language with topics from the political domain. The first dataset is a collection of **110** articles from Wikipedia, while the second dataset includes **10** articles from Aljazeera.net. Each article in the Wikipedia dataset, has an already constructed infobox. The infoboxes that our prototype produces for the same articles are evaluated against the already constructed infoboxes. On the other hand, we manually constructed infoboxes for the Aljazeera dataset. Again, the infoboxes generated by our prototype for the same articles from Aljazeera are compared with the manually constructed ones in order to evaluate the quality of results. Our prototype uses the first few lines of the text in each article in order to extract the relevant information.

5.2 Experiments on Dataset 1: Wikipedia Articles

5.2.1 Basic information about the dataset

- 110 political character articles.
- These articles are not stub articles.
- For evaluation purposes, the information and the infoboxes in the articles are removed.

Our system will create infoboxes for these articles using only the first 1000 character of the original document ignoring the rest of document. The created infoboxes by our system will be compared to the existing infoboxes that are originally in the articles.

Below is a sample of collected dataset:

Table (5.1): Wikipedia Dataset Sample

#	Name	Full Name	Birth Location	Birth Date
1.	سلمان بن عبد العزيز آل سعود	سلمان بن عبد العزيز بن عبد الرحمن بن فيصل بن تركي آل سعود	الرياض	5 شوال 1354 هـ/31 ديسمبر 1935
2.	عبد الفتاح السيسي	عبد الفتاح سعيد حسين خليل السيسي	القاهرة	19 نوفمبر 1954
3.	عبد العزيز بوتفليقة	عبد العزيز بوتفليقة	وجدة	2 مارس 1937
4.	محمد السادس بن الحسن	هو محمد بن الحسن بن محمد بن يوسف بن الحسن بن محمد بن عبد الرحمن بن هشام بن محمد بن عبد الله الخطيب بن إسماعيل بن مولاي علي الشريف العلوي	-	21 أغسطس 1963
5.	عبد الله الثاني بن الحسين	عبد الله الثاني بن الحسين بن طلال بن عبدالله بن حسين الهاشمي	-	30 يناير 1962
6.	خليفة بن زايد آل نهيان	الشيخ خليفة بن زايد بن سلطان بن زايد بن خليفة بن شخبوط بن ذياب آل نهيان الفلاحي	أبو ظبي	25 يناير 1948
7.	عبد ربه منصور هادي	عبد ربه منصور هادي	قرية ذكين	1 سبتمبر 1945
8.	قابوس بن سعيد	قابوس بن سعيد بن تيمور بن فيصل بن تركي بن سعيد بن	سلطنة عُمان	18 نوفمبر 1940م

#	Name	Full Name	Birth Location	Birth Date
		سلطان بن أحمد بن سعيد آل بو سعيدي		
9.	عمر البشير	عمر حسن أحمد البشير	قرية صغيرة تسمى حوش بانقا بريفي شندي ينتمي لقبيلة البديرية الدهمشية	1 يناير 1944
10	فؤاد معصوم	محمد فؤاد معصوم هورام	كويسنجق في شمال العراق	1938م

5.2.2 Extracting Locations

5.2.2.1 Birth Location

Input: The Testing Data-set (TD) experiment is a sum of **110** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of named entities representing birth location.

- **Experiment Settings**

As we mentioned in chapter 3, we use Lucene as search engine with settings below:

- Standard Analyzer
- Default Operator (AND)

We conducted 1-4 experiments during development phase, we use a data "development data set" Arabic Wikipedia articles from a total of **110** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**
  1- **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all of articles text to sentences then calculated distance between search keyword (ولد ، نشأ) and location entities then show the minimum distance as candidate, this way wasn’t accurate because location entity can be in the same sentence of search keyword.
  2- **Experiment 2:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then calculated distance between search keyword (ولد ، نشأ) and location entities then show the minimum distance as candidate.
- **Testing Phase Experiments**
  1- **Experiment 3:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment the sentence that contains the search keyword (ولد ، نشأ) to words then calculated distance between search keyword (ولد ، نشأ) and location entities then show the minimum distance as candidate.
  2- **Experiment 4:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we applied all of our experiments using ATKS to insure and increase results accuracy.

The resulted outcomes of development phase’s experiment 1 showed that the birth locations are not detected in most of articles. But experiment 2 showed that the birth locations are detected more accurate than experiment 1. Therefore, we applied the same settings of experiment 2 on experiment 3 which was conducted on testing data-set with more performance because we search for location entity in a short paragraph that contains keyword instead of all article.

Experiment 3: The detect birth location is 32 of 80 (80 is related to articles that have (ولد،نشأ) keywords) articles size and the low detected locations is related to NER tool (GATE), Table 1,2 shown samples of detection results:

Table (5.2): Experiment 3-Success Detection Samples

Article Name	Snippet	Result
بشار الأسد	ولد بشار الأسد بدمشق بسوريا	سوريا
تميم بن حمد بن خليفة آل ثاني	ولد في الدوحة	الدوحة
محمد نجيب	ولد محمد نجيب في السودان	السودان
خالد مشعل	ولد في سلواد قضاء رام الله بفلسطين	فلسطين
سعد الدين الحريري	ولد في السعودية	السعودية

Table (5.3): Experiment 3-Failed Detection Samples

Article Name	Snippet	Result
خليفة بن زايد آل نهيان	ولد في مدينة أبو ظبي	مدينة
عمر البشير	ولد في قرية صغيرة تسمى حوش بنقا	-
صباح الأحمد الجابر الصباح	ولد في مدينة الجهراء	-
محمد ولد عبد العزيز	ولد بدمينة اكجوجت	مدينة
أمين الحسيني	ولد في مدينة القدس	مدينة

- **Basic Results:**

We evaluate the quality of the generated detected birth locations by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = 32 / 80 = 40%

As we show above, we have a very low percentage of detection so, to enhance the accuracy, we use ATKS instead of GATE in Experiment 4:

The detect birth location is 65 of 80 (80 is related to articles that have (ولد،نشأ) keywords) articles size Table 5.4 shown samples of detection results:

Table (5.4): Experiment 4-Success Detection Samples

Article Name	Snippet	Result
بشار الأسد	ولد بشار الأسد بدمشق بسوريا	دمشق
تميم بن حمد بن خليفة آل ثاني	ولد في الدوحة	الدوحة
محمد نجيب	ولد محمد نجيب في السودان	الخرطوم
خالد مشعل	ولد في سلواد قضاء رام الله بفلسطين	سلواد
سعد الدين الحريري	ولد في السعودية	السعودية

Table (5.5): Experiment 4-Success Detection Samples that Failed in Gate

Article Name	Snippet	Result
خليفة بن زايد آل نهيان	ولد في مدينة أبو ظبي	مدينة
صباح الأحمد الجابر الصباح	ولد في مدينة الجهراء	الجهراء
محمد ولد عبد العزيز	ولد بدمينة اكجوجت	اكجوجت
أمين الحسيني	ولد في مدينة القدس	القدس

Table (5.6): Experiment 4-Failed Detection Samples

Article Name	**Snippet**	**Result**
خليفة بن زايد آل نهيان	ولد في مدينة أبو ظبي	-
عمر البشير	ولد في قرية صغيرة تسمى حوش بنقا	ماليزيا
عبد الله الثاني بن الحسين	-	نوفمبر
صدام حسين	ولد في مدينة العوجة	السبعينات

- **Improved Results**

We evaluate the quality of the generated detected birth locations by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (65 * 100) / 80 = 81.25%

5.2.2.2 Death Location

Input: The Testing Data-set (TD) experiment is a sum of **110** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of entities locations.

- **Experiment Settings**

As we mentioned in section 3, we use Lucene as search engine with settings below:

- Standard Analyzer
- Default Operator (AND)

In order to identify theses parameters, we conducted 1- 4 experiments during development phase, we use a data "development data set "Arabic Wikipedia articles from a total of **110** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**
    1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all of articles text to sentences then calculate distance between search keyword (مات ، توفي) and location entities then show the minimum distance as candidate, this way wasn't accurate because location entity can be in the same sentence of search keyword.
    2. **Experiment 2:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then calculate distance between search keyword (مات ، توفي) and location entities then show the minimum distance as candidate.
- **Testing Phase Experiments**
    1. **Experiment 3:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment the sentence that contains the search keyword
       (مات ، توفي) to words then calculate distance between search keyword (مات)
       توفي ، and location entities then show the minimum distance as candidate.
    2. **Experiment 4:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we apply all of our experiments using ATKS to insure and increase results accuracy.

The resulted outcomes of development phase's experiment 1 showed that the death locations are not detected in most of articles. But experiment 2 showed that the death locations are detected more accurate than experiment 1. Therefore, we applied the same settings of experiment 2 on experiment 3 which was conducted on testing data-set with more performance because we search for location entity in a short paragraph that contains keyword instead of all article.

The detected death location is 7 of 9 (9 is related to articles that have death location) articles size, to enhance the accuracy, Table 11,12 shown samples of detection results the full list of detected location extracted, we use ATKS instead of GATE in Experiment 4:

Table (5.7): Experiment 4-Success Detection Samples

Article Name	Snippet	Result
أمين الحسيني	توفي ببيروت	بيروت
عبد العزيز الرنتيسي	توفي بمدينة غزة	غزة
نزار ريان	في جباليا بالقرب من غزة	غزة

Table (5.8): Experiment 4-Failed Detection Samples

Article Name	Snippet	Result
فتحي الشقاقي	أغتيل بواسطة الموساد الصهيوني في مالطا.	-
الشريف حسين	توفي سنة 1350 هـ - 1931 م، فحمل إلى القدس ودفن فيها.	-

- **Results**

We evaluate the quality of the generated detected death locations by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (7 / 9) * 100 = 77%

We selected the articles that have the common targeted named entities but due to the lack of a sufficient number of articles that contain death location we ignored experiment result from the total result.

5.2.3 Extracting Full Name

Input: The Testing Data-set (TD) experiment is a sum of **110** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of entities full name.

- **Experiment Settings**

In order to identify theses parameters, we conducted 1-2 experiments during development phase, we use a data “development data set “Arabic Wikipedia articles from a total of 110 Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then detect named entities (Person) and calculate distance between these entities then show candidates.

- **Testing Phase Experiments**

1. **Experiment 2:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we apply experiment using ATKS.

The resulted outcomes of development phase’s experiment 1 showed that the named entities (Person) are not detected in most of articles. But experiment 2 showed that using ATKS in detection is more accurate than experiment 1.

The detected full name is 88 of 110 articles size, Table 5.9, and Table 5.10 show samples of detection results the full list of detected full name.

Table (5.9): Experiment 2-Success Detection Samples

Article Name	**Expected Result**	**Detected Result**
قابوس بن سعيد	قابوس بن سعيد بن تيمور بن فيصل بن تركي بن سعيد بن سلطان بن أحمد بن سعيد آل بو سعيدي	قابوس بن سعيد بن تيمور بن فيصل بن تركي بن سعيد بن سلطان بن أحمد بن سعيد البوسعيدي
عبد الفتاح السيسي	عبد الفتاح سعيد حسين خليل السيسي	عبد الفتاح سعيد حسين خليل السيسي
محمد السادس بن الحسن	هو محمد بن الحسن بن محمد بن يوسف بن الحسن بن محمد بن عبد الرحمن بن هشام بن محمد بن عبد	محمد بن الحسن بن محمد بن يوسف بن الحسن بن محمد بن عبد الرحمن بن هشام بن محمد بن عبد الله

Article Name	Expected Result	Detected Result
	الله الخطيب بن إسماعيل بن مولاي علي الشريف العلوي	الخطيب بن إسماعيل بن مولاي علي
عبد الله الثاني بن الحسين	عبد الله الثاني بن الحسين بن طلال بن عبدالله بن حسين الهاشمي	عبد الله الثاني بن الحسين بن طلال بن عبدالله بن حسين الهاشمي
خليفة بن زايد آل نهيان	الشيخ خليفة بن زايد بن سلطان بن زايد بن خليفة بن شخبوط بن ذياب آل نهيان الفلاحي	خليفة بن زايد بن سلطان بن زايد بن خليفة بن شخبوط بن ذياب آل نهيان

Table (5.10): Experiment 2-Failed Detection Samples

Article Name	Expected Result	Detected Result
حكم بلعاوي	حكم بلعاوي	بلعاوي
حافظ الاسد	حافظ الأسد	1966-1972
خليل الوزير	خليل الوزير	خليل

- **Results**

We evaluate the quality of the generated detected Full Name by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (88 / 110) * 100 = 80.9%

5.2.4 Extracting Dates

5.2.4.1 Birth Date

Input: The Testing Data-set (TD) experiment is a sum of **110** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of birth, death Dates. Our algorithm returns the best candidate of birth, death that provided from voting.

- **Experiment Settings**

We use LingPipe as a rule based named entity detector with settings below:

a. Date Regular Expression (/(\d{1,2})(\s)(.{4,15})(\s)(\d{2,4})/g)

In order to identify theses parameters, we conducted 1 experiment during development phase, we use a data “development data set “Arabic Wikipedia articles from a total of 110 Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We write a regular expression that captures dates entities from articles then show birth date.

- **Testing Phase Experiments**

The resulted outcomes of development phase’s experiment 1 showed that the birth locations are detected in most of articles.

Experiment 1: The detect birth date is **69** of **105** articles (105 is related to articles that have birthdate) Table 5.11 and Table 5.12 shown samples of detection results the full list of detected birth date.

Table (5.11): Experiment 1-Success Detection Samples

Article Name	Expected Result	Detected Result
سلمان بن عبد العزيز آل سعود	5 شوال 1354 هـ	5 شوال 1354 هـ
عبد الفتاح السيسي	19 نوفمبر 1954	19 نوفمبر 1954
محمد السادس بن الحسن	21 أغسطس 1963	21 أغسطس 1963
عبد الله الثاني بن الحسين	30 يناير 1962	30 يناير 1962
خليفة بن زايد آل نهيان	25 يناير 1948	25 يناير 1948

Table (5.12): Experiment 1-Failed Detection Sample

Article Name	Expected Result	Detected Result
فؤاد معصوم	1938م	-
تمام سلام	1945	-
مصطفى البرغوثي	1954 عام	-

- **Results**

We evaluate the quality of the generated detected birth date by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (69 / 105) * 100 = 65.7%

5.2.4.2 Death Date

Input: The Testing Data-set (TD) experiment is a sum of **110** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of birth, death Dates. Our algorithm returns the best candidate of birth, death that provided from voting.

- **Experiment Settings**

We use LingPipe as a rule based named entity detector with settings below:

1- Date Regular Expression (/(\d{1,2})(\s)(.{4,15})(\s)(\d{2,4})/g)

In order to identify theses parameters, we conducted 1 experiment during development phase, we use a data “development data set “Arabic Wikipedia articles from a total of **110** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1- **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We write a regular expression that captures dates entities from articles then show death date.

- **Testing Phase Experiments**

The resulted outcomes of development phase’s experiment 1 showed that the birth locations are detected in most of articles.

Experiment 1: The detect birth date is **35** of **40** articles (40 is related to articles that have death date) size the full list of detected. Table 5.13 and Table 5.14 show samples of detection results the full list of detected death date.

Table (5.13): Experiment 1-Success Detection Samples

Article Name	Expected Result	Detected Result
محمد أنور السادات	6 أكتوبر 1981م	6 أكتوبر 1981م
جمال عبد الناصر	28 سبتمبر 1970	28 سبتمبر 1970
محمد نجيب	28 أغسطس 1984	28 أغسطس 1984
أمين الحسيني	4 يوليو 1974	4 يوليو 1974
أحمد ياسين	22 مارس 2004	22 مارس 2004

Table (5.14): Experiment 2-Failed Detection Samples

Article Name	Expected Result	Detected Result
نمر حماد	2016	-
فتحي الشقاقي	1995	-
الشريف حسين	1854م	-

- **Results**

We evaluate the quality of the generated detected death date by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (35 / 40) * 100 = 87.5%

5.3 Experiments on Dataset 2: Aljazeera articles

5.3.1. Basic information about the dataset

- We collect 10 political characters' articles from Aljazeera that are also articles in Wikipedia, we manually collect birth, death date, location and full name from these articles.
- We run our system on the articles from Aljazeera in order to build infoboxes.

- We then compare the results we obtain from our system with the existing infoboxes in the corresponding articles in Wikipedia.

Below is a collected dataset:

Table (5.15): Aljazeera Dataset Sample

#	Name	Full Name	Birth Location	Birth Date
1.	محمد نجيب	محمد نجيب يوسف	ساقية أبو العلا بالخرطوم	20فبراير 1901
2.	الملك فاروق	-	القاهرة	11فبراير 1920
3.	أنور السادات	محمد أنور السادات	ميت أبو الكوم التابعة لمحافظة المنوفية	25ديسمبر 1918
4.	جمال عبد الناصر	-	حي باكوس بالإسكندرية	15يناير 1918
5.	الحسين بن طلال	-	عمان	14نوفمبر 1935
6.	خليفة بن زايد آل نهيان	-	قلعة المويجعي بمدينة العين في إمارة أبو ظبي	1948
7.	زايد بن سلطان ال نهيان	-	أبوظبي	1918
8.	قابوس بن سعيد	-	مدينة صلالة بمحافظة ظفار	18نوفمبر 1940
9.	بسام الصالحي	بسام أحمد عمر الصالحي	رام الله	16يناير 1960
10.	فؤاد معصوم	محمد فؤاد معصوم هورامي	أربيل كردستان العراق	1 يناير 1938

5.3.2 Extracting Locations

5.3.2.1 Birth Location

Input: The Testing Data-set (TD) experiment is a sum of **10** articles. These are sampled from a total of Aljazeera articles data-set.

Output: A set of named entities representing birth location.

- **Experiment Settings**

As we mentioned in chapter 3, we use Lucene as search engine with settings below:

  - Standard Analyzer
  - Default Operator (AND)

We conducted 1-4 experiments during development phase, we use a data "development data set" Arabic Wikipedia articles from a total of **10** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1- **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all of articles text to sentences then calculated distance between search keyword (ولد ، نشأ) and location entities then show the minimum distance as candidate, this way wasn't accurate because location entity can be in the same sentence of search keyword.

2- **Experiment 2:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then calculated distance between search keyword (ولد ، نشأ) and location entities then show the minimum distance as candidate.

- **Testing Phase Experiments**

3- **Experiment 3:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we applied all of our experiments using ATKS to insure and increase results accuracy.

The resulted outcomes of development phase's experiment 1 showed that the birth locations are not detected in most of articles. But experiment 2 showed that the birth locations are detected more accurate than experiment 1. Therefore, we applied the same settings of experiment 2 on experiment 3 which was conducted on testing data-set with more performance because we search for location entity in a short paragraph that contains keyword instead of all article.

Experiment 3: The detect birth location is **10** of **10** articles size, Table 1,2 shown samples of detection results:

Table (5.16): Experiment 3-Success Detection Samples

Article Name	**Snippet**	**Result**
محمد نجيب	ساقية أبو العلا بالخرطوم	الخرطوم
الملك فاروق	القاهرة	القاهرة
أنور السادات	ميت أبو الكوم التابعة لمحافظة المنوفية	المنوفية
جمال عبد الناصر	حي باكوس بالإسكندرية	الاسكندرية
الحسين بن طلال	عمان	عمان

- **Results:**

We evaluate the quality of the generated detected birth locations by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = 10 / 10 = 100%

5.3.2.2 Death Location

Input: The Testing Data-set (TD) experiment is a sum of **10** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of entities locations.

- **Experiment Settings**

As we mentioned in section 3, we use Lucene as search engine with settings below:

- Standard Analyzer
- Default Operator (AND)

In order to identify theses parameters, we conducted 1- 4 experiments during development phase, we use a data "development data set "Arabic Wikipedia articles from a total of **10** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**
  1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all of articles text to sentences then calculate distance between search keyword (مات ، توفي) and location entities then show the minimum distance as candidate, this way wasn't accurate because location entity can be in the same sentence of search keyword.
  2. **Experiment 2:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then calculate distance between search keyword (مات ، توفي) and location entities then show the minimum distance as candidate.
- **Testing Phase Experiments**
  3. **Experiment 3:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we apply all of our experiments using ATKS to insure and increase results accuracy.

The resulted outcomes of development phase's experiment 1 showed that the death locations are not detected in most of articles. But experiment 2 showed that the death locations are detected more accurate than experiment 1. Therefore, we applied the same settings of experiment 2 on experiment 3 which was conducted on testing data-set with more performance because we search for location entity in a short paragraph that contains keyword instead of all article.

The detected death location is **0** of **1** (1 is related to articles that have death location) articles size, Table 11,12 shown samples of detection results the full list of detected location extracted, we use ATKS instead of GATE in Experiment 3:

Table (5.17): Experiment 3- Failed Detection Samples

Article Name	Snippet	Result
الملك فاروق	توفي بإيطاليا	-

- **Results**

We evaluate the quality of the generated detected death locations by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (0 / 1) * 100 = 0%

We selected the articles that have the common targeted named entities but due to the lack of a sufficient number of articles that contain death location we ignored experiment result from the total result.

5.3.3 Extracting Full Name

Input: The Testing Data-set (TD) experiment is a sum of **10** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of entities full name.

- **Experiment Settings**

In order to identify theses parameters, we conducted 1-2 experiments during development phase, we use a data “development data set “Arabic Wikipedia articles from a total of 110 Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We segment all articles text to words then detect named entities (Person) and calculate distance between these entities then show candidates.

- **Testing Phase Experiments**

2. **Experiment 2:** we use GATE tool in all of the above experiments for detecting named entities, in this experiment we apply experiment using ATKS.

The resulted outcomes of development phase's experiment 1 showed that the named entities (Person) are not detected in most of articles. But experiment 2 showed that using ATKS in detection is more accurate than experiment 1.

The detected full name is 4 of 4 articles size (4 is related to articles that have death location) articles size, Table 13,14 shown samples of detection results the full list of detected full name.

Table (5.18): Experiment 2-Success Detection Samples

Article Name	Expected Result	Detected Result
محمد نجيب	محمد يوسف نجيب	محمد يوسف نجيب
بسام الصالحي	بسام أحمد عمر الصالحي	بسام أحمد عمر الصالحي
فؤاد معصوم	محمد فؤاد معصوم هورامي	محمد فؤاد معصوم هورامي
أنور السادات	محمد أنور السادات	محمد أنور السادات

- **Results**

We evaluate the quality of the generated detected Full Name by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (4 / 4) * 100 = 100%

5.3.4 Extracting Dates

5.3.4.1. Birth Date

Input: The Testing Data-set (TD) experiment is a sum of **10** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of birth, death Dates. Our algorithm returns the best candidate of birth, death that provided from voting.

- **Experiment Settings**

We use LingPipe as a rule based named entity detector with settings below:

b. Date Regular Expression (/(\d{1,2})(\s)(.{4,15})(\s)(\d{2,4})/g)

In order to identify theses parameters, we conducted 1 experiment during development phase, we use a data “development data set “Arabic Wikipedia articles from a total of 10 Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

1. **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We write a regular expression that captures dates entities from articles then show birth date.

- **Testing Phase Experiments**

The resulted outcomes of development phase’s experiment 1 showed that the birth locations are detected in most of articles.

2. **Experiment 2:** The detect birth date is **8** of **10** articles Table 15,16 shown samples of detection results the full list of detected full name.

Table (5.19): Experiment 1-Success Detection Samples

Article Name	Expected Result	Detected Result
محمد نجيب	20فبراير 1901	20فبراير 1901
الملك فاروق	11فبراير 1920	11فبراير 1920
أنور السادات	25ديسمبر 1918	25ديسمبر 1918
جمال عبد الناصر	15يناير 1918	15يناير 1918

Table (5.20): Experiment 1-Failed Detection Sample

Article Name	Expected Result	Detected Result
خليفة بن زايد آل نهيان	1948	-
زايد بن سلطان ال نهيان	1918	-

- **Results**

We evaluate the quality of the generated detected birth date by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (8 / 10) * 100 = 80%

5.3.4.2. Death Date

Input: The Testing Data-set (TD) experiment is a sum of **10** articles. These are sampled from a total of Arabic Wikipedia articles data-set.

Output: A set of birth, death Dates. Our algorithm returns the best candidate of birth, death that provided from voting.

- **Experiment Settings**

We use LingPipe as a rule based named entity detector with settings below:

1- Date Regular Expression (/(\d{1,2})(\s)(.{4,15})(\s)(\d{2,4})/g)

In order to identify theses parameters, we conducted 1 experiment during development phase, we use a data "development data set "Arabic Wikipedia articles from a total of **10** Arabic articles from political domain. We describe the experiments we conducted during development and testing phase.

- **Development Phase Experiments**

2- **Experiment 1:** In this experiment, we calculated accuracy of articles from a data set of some articles. We write a regular expression that captures dates entities from articles then show death date.

- **Testing Phase Experiments**

The resulted outcomes of development phase's experiment 1 showed that the birth locations are detected in most of articles.

Experiment 1: The detect birth date is **5** of **6** articles (6 is related to articles that have death date). Table 5.21 and Table 5.22 show samples of detection results the full list of detected death date.

Table (5.21): Experiment 1-Success Detection Samples

Article Name	Expected Result	Detected Result
محمد نجيب	6 أكتوبر 1981م	6 أكتوبر 1981م
الملك فاروق	28 سبتمبر 1970	28 سبتمبر 1970
أنور السادات	28 أغسطس 1984	28 أغسطس 1984
جمال عبد الناصر	4 يوليو 1974	4 يوليو 1974

Table (5.22): Experiment 2-Failed Detection Samples

Article Name	Expected Result	Detected Result
زايد بن سلطان ال نهيان	2نوفمبر/تشرين الثاني 2004	-

- **Results**

We evaluate the quality of the generated detected death date by accuracy equation:

Accuracy = Detected Articles / Total Numbers of Articles

Accuracy = (5 / 6) * 100 = 90%

5.4 Discussion of Results

We calculate the overall accuracy of each extraction algorithm in our prototype as follows (see Table 5.23 below):

Accuracy = number of articles in both Wikipedia dataset and Aljazeera dataset from which the entities were correctly extracted by our prototype / total number of articles in both Wikipedia dataset and Aljazeera dataset that actually contain the entities

5.4.1. Birth Location

Accuracy = (((65 + 10) / 90) *100) = 83.3%

5.4.2. Full Name

Accuracy = (((88 + 4) / 114) *100) = 80.7%

5.4.3. Birth Date

$$\text{Accuracy} = (((69 + 8) / 115) *100) = 66.9\%$$

5.4.4. Death Date

$$\text{Accuracy} = (((35 + 5) / 46) *100)) = 86.9\%$$

Table (5.23): Final Result

Extraction of	Location		Full Name	Dates	
	Birth	Death		Birth	Death
Accuracy	83.3%	-	80.7%	66.9%	86.9%
Average	83.3%		80.7%	76.9%	
Overall Average	80.3%				

As shown in table above, the results are very satisfactory compared to the complexity of working with Arabic language, the lack of available tools, as well as the weakness of the content on the internet. Achieving these results came after doing many experiments with the help of tools such as GATE, ATKS, Lucene, and LingPipe.

We did many experiments for each algorithm in order to improve its accuracy. In the future we will work on enhancing the accuracy of our system, try to increase number of patterns in birth and death date detection algorithm and applying voting algorithm.

Chapter 6

Conclusions and Future Work

Chapter 6

Conclusions and Future Work

In this research, we studied the possibility of supporting the content of Arabic Wikipedia articles by developing a **Semi-Automatic Method for Infoboxes Extraction.**

We studied the statistics of Arabic content over the internet in addition to Knowing Wikipedia articles structure, stub articles, and categories. We selected a politics domain as a domain of interest and collected a data set of many politics characters.

We developed a method for constructing infobox, this method has a five different algorithms for detecting Full Name, Birth Location, Death Location, Birth Date and Death date. We faced some of challenges and limitation that affect our system accuracy such as lack of tools and Arabic articles on the internet.

We conducted several experiments to evaluate our application and its different algorithms. The End-to-End evaluation of application shows that it has an accuracy of 80.3%. This is very satisfactory results compared to lack of tools and researches in Arabic language.

At future work we will work on enhancing the accuracy of our system, try to increase number of patterns in birth and death date detection algorithm, applying voting algorithm and extending our work with a feature of generate new textual content for stub articles in Arabic Wikipedia.

During the work on this thesis, we started to work on "voting algorithm". By adding this algorithm to our prototype, we aim to improve our results. We have conducted preliminary work on this algorithm, but it has not been evaluated yet. This is left as a future work, but we found it would be important to present out preliminary work on this algorithm in the next few lines.

In brief, in order to increase our results accuracy, we collect more than one article for each person from many resources. We apply our algorithm for the collected articles and show the common result as a final result.

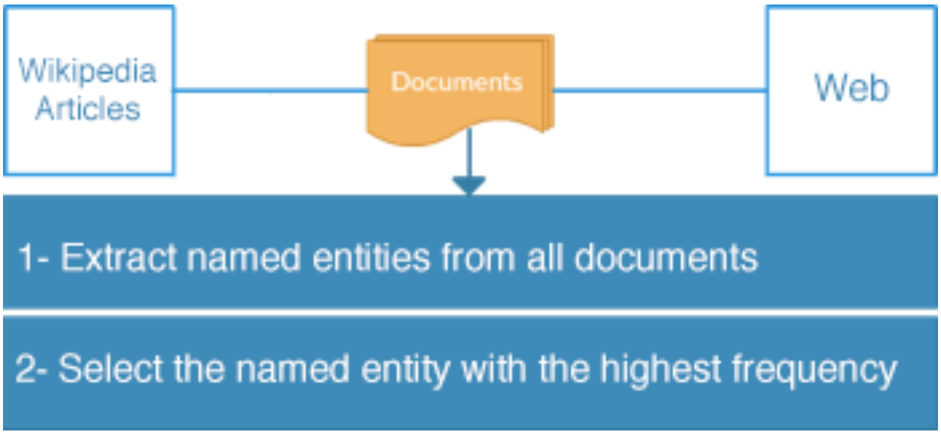

Wikipedia Articles
Documents
Web
1- Extract named entities from all documents
2- Select the named entity with the highest frequency

Figure (6.1): Voting Process

Below is an example that explain the steps of algorithm:

- " ولد في القاهرة [1] لأسرة فلسطينية.[5] أبوه عبد الرؤوف عرفات القدوة الحسيني من غزة، وجدته مصرية .وكان أبوه يعمل في تجارة الأقمشة في حي السكاكيني . وكان الولد السادس لأسرة فلسطينية تتكون من سبعة أفراد. ولد هو وأخوه الصغير فتحي في القاهرة. ونسبه من جهة أمه يتفرع من عائلة الحسيني، التي تعتبر من الأسر المقدسية المعروفة والتي برز بعض أفرادها في التاريخ الوطني الفلسطيني[6] .قضى عرفات مراحل طفولته ومرحلة شبابه الأولى في القاهرة .وقد توفيت والدته زهوة أبو السعود عندما كان في الرابعة من عمره بسبب قصور كلوي[6].[4]"
- " الشهيد الخالد ياسر عرفات 'أبو عمار ' سيرة ذاتية .. مسيرة شعب أبو عمار (4 أغسطس 1929 - 11 نوفمبر 2004)، هو محمد عبد الرحمن عبد الرؤوف عرفات القدوة الحسيني و كنيته (أبو عمار)، رئيس السلطة الوطنية الفلسطينية المنتخب في عام 1996. ترأس منظمة التحرير الفلسطينية سنة 1969، وهو قائد فتح أكبر الحركات داخل المنظمة. فاز مع إسحاق رابين بجائزة نوبل للسلام سنة 1994. السيرة الذاتية:اسمه محمد عبد الرحمن، وهو اسم مركب واسم أبيه هو عبد الرؤوف واسم جده عرفات واسم عائلته القدوة من عشيرة الحسيني وهو واحد من سبعة إخوة ولدوا لتاجر فلسطيني ولد في مدينة القدس في 4 أغسطس/ آب 1929."
- "لا يعرف على وجه اليقين مكان ولادة محمد عبد الرؤوف القدوة الحسيني الذي اشتهر فيما بعد باسم ياسر عرفات أو أبو عمار، والأغلب أنه ولد في القاهرة في الرابع والعشرين من أغسطس/ آب عام 1929، وهو الابن السادس لأب كان يعمل في التجارة، وهاجر إلى القاهرة عام 1927 وعاش في حي السكاكيني. وعندما توفيت والدته وهو في الرابعة من عمره أرسله والده إلى القدس، وهناك بدأ وعيه يتفتح على أحداث ثورة 1936. "

1. After applying our method on three articles above, the following named entities are returned for the three articles respectively:
    a) القاهرة
    b) القدس
    c) القاهرة
2. The selected name entity is "القاهرة" since it has the highest frequency

The Reference List

The Reference List

Al Zamil, M. G., & Al-Radaideh, Q. (2014). Automatic extraction of ontological relations from Arabic text. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 462-472.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data The semantic web* (pp. 722-735): Springer.

Apache .(2016) . *Lucene* ,Retrieved: 22 July 2016 , from https://lucene.apache.org/

Apache .(2017) . *lucene* , Retrieved 11-Feb-2017 ,from http://lucene.apache.org/core/

Andreas .(2015). *Information Retrieval Challenge* , Retrieved 2 May 2016 , from https://artios.io/information-retrieval-challenges#.Vg0VbbT92fQ

Banerjee, S., & Mitra, P. (2016). *WikiWrite: Generating Wikipedia Articles Automatically.*

Boujelben, I., Jamoussi, S., & Hamadou, A. B. (2014). A hybrid method for extracting relations between Arabic named entities. *Journal of King Saud University-Computer and Information Sciences*, *26*(4), 425-440.

Cheddadi, A. (2014). *Three-levels Approach for Arabic Question Answering Systems*. Ecole Mohammadia d'Ingénieurs.

Elkateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (24-25-26 MAY 2006). *Building a wordnet for arabic*. Paper presented at the Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006).

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Alkhalifa, M. (23-Oct-2006 - 23-Oct-2006). *Arabic WordNet and the challenges of Arabic.* Paper presented at the Proceedings of Arabic NLP/MT Conference, the British Computer Society,London, UK.

Elsebai, A., & Meziane, F. (25 Apr - 27 Apr 2011). *Extracting person names from Arabic newspapers*. Paper presented at the Innovations in Information Technology (IIT), 2011 International Conference on, the Faculty of Information Technology (FIT), United Arab Emirates University (UAEU)

GATE .(1995) . *Named Entity Recognition* , Retrieved 11 August 2016 , from https://gate.ac.uk/

Google .(2008). *wikixmlj* , Retrieved 11 November 2015 , from https://code.google.com/archive/p/wikixmlj/

Fader, A., Soderland, S., & Etzioni, O. (October 25 - 27, 2008). *Identifying relations for open information extraction*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing. University of Pennsylvania

Fellbaum, C. (1998a). *A semantic network of English verbs*. WordNet: An electronic lexical database, 3, 153-178.

Fellbaum, C. (1998b). *WordNet*: Wiley Online Library.

Hammo, B., Abu-Salem, H., & Lytinen, S. (29 June 2005). *QARAB: A question answering system to support the Arabic language*. Paper presented at the Proceedings of the ACL-02 workshop on Computational approaches to semitic languages. University of Michigan

Fellbaum, Christiane .(2005). *WordNet* , Retrieved 16 Feb 2016 , from https://wordnet.princeton.edu

Hirschman, L., & Gaizauskas, R. (2001). Natural language question answering: the view from here. *natural language engineering*, *7*(04), 275-300.

JWPL -Java Wikipedia Library. (2016). Retrieved 11-Feb-2017 ,from https://dkpro.github.io/dkpro-jwpl/

Kim, M.-k., & Kim, H.-J. (2-4 sept 2008). *Design of question answering system with automated question generation*. Paper presented at the Networked Computing and Advanced Information Management, 2008. NCM'08. Fourth International Conference on.

Lange, D., Böhm, C., & Naumann, F. (2010). *Extracting structured information from Wikipedia articles to populate infoboxes*. Paper presented at the Proceedings of the 19th ACM international conference on Information and knowledge management.

LingPipe.(2011) . Named Entity Recognition, Retrieved 11-Feb-2017 ,from http://alias-i.com/lingpipe/demos/tutorial/ne/read-me.html

Microsoft .(2013) . *ATKS* , Retrieved 25 November 2016 , from https://www.microsoft.com/en-us/research/project/arabic-toolkit-service-atks/

NetBeans.(2017) . *Netbeans IDE* , Retrieved 11-Feb-2017 ,from https://netbeans.org/

Oudah, M., & Shaalan, K. F. (2012). *A Pipeline Arabic Named Entity Recognition using a Hybrid Approach*. Paper presented at the COLING.

Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M. A., Pease, A. (2008). *Arabic wordnet: Current state and future extensions.* Paper presented at the Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary.

Rothfels, J., Saeta, B., & Topalovic, E. (2011). *A recommendation engine for Wikipedia articles based on constrained training data*: PDF Open access.

Saste, R. P., & Patil, S. S. (2014). *Extraction of incremental information using query evaluator. Paper* presented at the Networks & Soft Computing (ICNSC), 2014 First International Conference on.

Sauper, C., & Barzilay, R. (2009). *Automatically generating wikipedia articles: A structure-aware approach*. Paper presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1.

Schmid, H. (2013). *Probabilistic part-of-speech tagging using decision trees*. Paper presented at the New methods in language processing.

Sriurai, W., Meesad, P., & Haruechaiyasak, C. (2009). *Recommending Related Articles in Wikipedia via a Topic-Based Model*. Paper presented at the IICS.

Sun, B. (2010). *Named entity recognition*: Evaluation of existing systems.

Tari, L., Tu, P. H., Hakenberg, J., Chen, Y., Son, T. C., Gonzalez, G., & Baral, C. (2012). Incremental information extraction using relational databases. *IEEE Transactions on Knowledge and Data Engineering, 24*(1), 86-99.

Torres, D., Skaf-Molli, H., Molli, P., & Díaz, A. (2013). *BlueFinder: recommending wikipedia links using DBpedia properties*. Paper presented at the Proceedings of the 5th Annual ACM Web Science Conference.

The Stanford Natural Language Processing Group. (2001). Stanford Word Segmenter , Retrieved 11-Feb-2017 ,from http://nlp.stanford.edu/software/segmenter.html

The Stanford Natural Language Processing Group .(2000) . Stanford Tokenizer , Retrieved 11-Feb-2017 , from http://nlp.stanford.edu/software/tokenizer.shtml

Wu, F., & Weld, D. S. (2008). *Automatically refining the wikipedia infobox ontology.* Paper presented at the Proceedings of the 17th international conference on World Wide Web.

Wikipedia. .(2015) . *قائمة اللغات حسب العدد الكلي للمتحدثين* , Retrieved 11 January 2016 , from https://ar.wikipedia.org/wiki/قائمة_اللغات_حسب_العدد_الكلي_للمتحدثين

Wikipedia. .(2006). List of Wikipedias , Retrieved 22 August 2015 , from https://en.wikipedia.org/wiki/List_of_Wikipedias.

Wikipedia. (2005) . Languages used on the Internet , Retrieved 11 January 2016 from https://en.wikipedia.org/wiki/Languages_used_on_the_Internet.

Yuncong, C., & Fung, P. (2010). *Unsupervised synthesis of multilingual Wikipedia articles*. Paper presented at the Proceedings of the 23rd International Conference on Computational Linguistics.

Zhang, K., Xiao, Y., Tong, H., Wang, H., & Wang, W. (2014). *WiiCluster: a Platform for Wikipedia Infobox Generation*. Paper presented at the Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management.

Appendices

Appendix 1: **Wikipedia Articles Dataset**

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
1.	سلمان بن عبد العزيز آل سعود	سلمان بن عبد العزيز بن عبد الرحمن بن فيصل بن تركي آل سعود	الرياض	-	5 شوال 1354 هـ/ 31 ديسمبر 1935	-
2.	عبد الفتاح السيسي	عبد الفتاح سعيد حسين خليل السيسي	القاهرة	-	19 نوفمبر 1954	-
3.	عبد العزيز بوتفليقة	عبد العزيز بوتفليقة	وجدة	-	2 مارس 1937	-
4.	محمد السادس بن الحسن	هو محمد بن الحسن بن محمد بن يوسف بن الحسن بن محمد بن عبد الرحمن بن هشام بن محمد بن عبد الله الخطيب بن إسماعيل بن مولاي علي الشريف العلوي	-	-	21 أغسطس 1963	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.5	عبد الله الثاني بن الحسين	عبد الله الثاني بن الحسين بن طلال بن عبدالله بن حسين الهاشمي	-	-	30 يناير (1962)	-
.6	خليفة بن زايد آل نهيان	الشيخ خليفة بن زايد بن سلطان بن زايد بن خليفة بن شخبوط بن ذياب آل نهيان الفلاحي	أبو ظبي	-	25 يناير 1948	-
.7	عبد ربه منصور هادي	عبد ربه منصور هادي	قرية ذكين	-	1 سبتمبر 1945	-
.8	قابوس بن سعيد	قابوس بن سعيد بن تيمور بن فيصل بن تركي بن سعيد بن سلطان بن أحمد بن سعيد آل بو سعيدي	سلطنة عُمان	-	18 نوفمبر 1940م	-
.9	عمر البشير	عمر حسن أحمد البشير	قرية صغيرة تسمى حوش بانقا بريفي شندي ينتمي لقبيلة البديرية الدهمشية	-	1 يناير 1944	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.10	فؤاد معصوم	محمد فؤاد معصوم هورام	كويسنجق في شمال العراق	-	1938م	-
.11	بشار الأسد	بشار حافظ الأسد	في دمشق بسوريا	-	11 سبتمبر 1965	-
.12	الباجي قائد السبسي	الباجي قائد السبسي محمد الباجي بن حسونة قائد السبسي	سيدي بوسعيد في تونس العاصمة	-	29 نوفمبر 1926	-
.13	تمام سلام	تمام صائب سلام	بيروت دراسة	-	1945	-
.14	صباح الأحمد الجابر الصباح	صباح الأحمد الجابر المبارك الصباح	مدينة الجهراء	-	16 يونيو 1929	-
.15	حمد بن عيسى بن سلمان آل خليفة	الملك حمد بن عيسى آل خليفة	-	-	28 يناير 1950	-
.16	تميم بن حمد بن خليفة آل ثاني	تميم بن حمد بن خليفة بن حمد بن عبدالله بن قاسم بن محمد آل ثاني	الدوحة	-	3 يونيو 1980	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.17	محمود عباس	محمود عباس ويكنى بـ أبو مازن	في مدينة صفد في فلسطين	-	26 مارس 1935	-
.18	حسن شيخ محمود	حسن شيخ محمود	-	-	29 نوفمبر 1955	-
.19	محمد ولد عبد العزيز	محمد ولد عبد العزيز ولد اعلية	بمدينة اكجوجت	-	20 ديسمبر 1956	-
.20	إكليل ظنين	إكليل ظنين (خليل ظنين)	-	-	14 أغسطس 1962	stub
.21	رفيق الحريري	رفيق الحريري رفيق بهاء الدين الحريري	صيدا في جنوب لبنان	-	1 نوفمبر 1944	14 فبراير 2005
.22	عبد الستار قاسم	عبد الستار توفيق قاسم الخضر	دير الغصون بطولكرم	-	-	-
.23	مروان البرغوثي	مروان البرغوثي	كوبر إلى الشمال الغربي من مدينة رام الله	-	6) يونيو 1958	-
.24	مصطفى البرغوثي	مصطفى كامل البرغوثي	مدينة القدس	-	عام 1954	-
.25	حسني مبارك	محمد حسني السيد مبارك	-	-	4 مايو 1928	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.26	محمد أنور السادات	محمد أنور محمد السادات	ميت أبو الكوم بمحافظة المنوفية	-	25 ديسمبر 1918م	6 أكتوبر 1981م)
.27	نبيل العربي	نبيل عبد الله العربي	-	-	15 مارس 1935	-
.28	جمال عبد الناصر	جمال عبد الناصر حسين	-رقم 12 شارع قنوات- بحي باكوس بالإسكندرية	-	15 يناير 1918	28 سبتمبر 1970
.29	محمد نجيب	حرب محمد نجيب	بالسودان	-	19 فبراير 1901	28 أغسطس 1984
.30	أحمد الشقيري	أحمد الشقيري	-	- في غور الأردن	1908	25 فبراير 1980
.31	أمين الحسيني	أمين الحسيني أو المفتي	في مدينة القدس	ببيروت	1895	4 يوليو 1974)
.32	حنان عشراوي	حنان داود خليل عشراوي	ولدت في نابلس	-	1946	-
.33	صائب عريقات	صائب محمد صالح عريقات	أبو ديس بالقدس	-	مواليد 28 أبريل 1955	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.34	أحمد سعدات	أحمد سعدات	بلدة دير طريف في الرملة	-	1953	-
.35	أحمد ياسين	أحمد إسماعيل ياسين	عريقة تسمى جورة عسقلان التابعة لقضاء مدينة المجدل	-	28) يونيو 1936	22 مارس 2004
.36	خالد مشعل	خالد عبد الرحيم إسماعيل عبد القادر مشعل	سلواد قضاء رام الله بفلسطين	-	28 مايو 1956	-
.37	عبد العزيز الرنتيسي	عبد العزيز علي عبد المجيد الحفيظ الرنتيسي	يبنا - تقع بين عسقلان ويافا	غزة	23 أكتوبر 1947	17 أبريل 2004
.38	نزار ريان	نزار عبد القادر محمد ريان العسقلاني	مخيم جباليا وتعود أصول أسرته إلى قرية نعليا إحدى قرى مدينة المجدل عسقلان	في جباليا بالقرب من غزة	6 مارس 1959	1 يناير 2009
.39	إسماعيل هنية	إسماعيل عبد السلام أحمد هنية	مخيم الشاطيء - عسقلان	-	1963	-
.40	محمود الزهار	محمود خالد الزهار	مدينة غزة في حي الزيتون	-	1945	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.41	سعيد صيام	سعيد محمد شعبان صيام	غزة ، الجورة ، عسقلان ، فلسطين	-	22 يوليو 1959	15 يناير 2009
.42	عبد الباري عطوان	عبد الباري عطوان	مخيم للاجئين بمدينة دير البلح في قطاع غزة	-	17 شباط سنة 1950.	-
.43	محمد نزال	محمد نزال "أبو براء"	قلقيلية	-	18 فبراير/ شباط 1963	-
.44	أمين الجميل	أمين الجميّل	-	-	22 يناير 1942	-
.45	صدام حسين	صدام حسين عبد المجيد التكريتي	العوجة التي تبعد 23 كم عن مدينة تكريت	-	28 أبريل 1937	30 ديسمبر 2006
.46	سعد الدين الحريري	سعد الدين الحريري	السعودية	-	18 أبريل 1970	-
.47	نبيه بري	نبيه بري	فريتاون عاصمة سيراليون	-	28 يناير 1938	-
.48	الحسين بن طلال	الحسين بن طلال بن عبداللہ بن حسين الهاشمي	عمّان	-	14 نوفمبر 1935	7 فبراير 1999

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.49	عزيز الدويك	عزيز سالم مرتضى الدويك	الخليل في مصر	-	12) يناير 1948	-
.50	نمر حماد	نمر حمّاد	الكويكات في عكا	بيروت	1941	2016
.51	عزام الأحمد	عزام الأحمد	قرية رمانة قضاء جنين	-	عام1947	-
.52	جمال الخضري	جمال ناجي الخضري	مواليد مدينة غزة	-	1955-	-
.53	رامي الحمد الله	رامي الحمد الله	عنبتا بمحافظة طولكرم	-	1958	-
.54	أسامة حمدان	أسامة حمدان	مدينة غزة	-	1965	-
.55	أحمد بحر	د. أحمد محمد عطية بحر	بغزة	-	1949	-
.56	روحي فتوح	روحي فتوح هو روحي أحمد محمد فتوح	برقة (فلسطين المحتلة	-	1948	-
.57	خالدة جرار	خالدة جرار	-	-	1963	stub
.58	ليلى خالد	ليلى خالد	حيفا	-	1944	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.59	أحمد الطيبي	أحمد كامل أحمد الطيبي	الطيبة لأب من يافا	-	19ديسمبر عام 1958م	-
.60	الملك فاروق	فاروق بن فؤاد بن إسماعيل بن إبراهيم باشا بن محمد علي باشا	-		11 فبراير 1920	18 مارس 1965) manual
.61	محمد بركة	محمد سعيد بركة	مدينة شفاعمرو	-	1955	-
.62	سليم الزعنون	سليم الزعنون	غزة	-	1933	stub
.63	ياسر عبد ربه	ياسر عبد ربه	يافا	-	1944	-
.64	أحمد قريع	أحمد علي قريع	أبوديس-القدس	-	1937	-
.65	بسام أبو شريف	بسام أبو شريف	القدس	-	1946	-
.66	عبد الحكيم عامر	محمد عبد الحكيم عامر	أسطال، مركز سمالوط بمحافظة المنيا	-	11 ديسمبر 1919	14 سبتمبر 1967

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.67	فتحي الشقاقي	فتحي إبراهيم عبد العزيز الشقاقي	مخيم الشاطئ بمدينة غزة	أغتيل بواسطة الموساد الصهيوني في مالطا.	4-1-1951	1995-
.68	جبريل الرجوب	جبريل الرجوب	ولد سنة 1953 في بلدة دورا،الخليل)	-	1953	-
.69	المنصف المرزوقي	محمد المنصف المرزوقي	قرمبالية		7 يوليو 1945	-
.70	الحبيب بورقيبة	الحبيب بورقيبة	الطرابلسية بمدينة المنستير الساحلية	-	3 أغسطس 1903	6 أبريل 2000
.71	أحمد بن بلة	أحمد بن بلّة	مغنية جنوب مدينة وهران بالغرب الجزائري	الجزائر	25 ديسمبر 1916	11 أبريل 2012

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.72	هواري بومدين	هواري بومدين واسمه الحقيقي محمد إبراهيم بوخروبة	في دوّار بني عدي (العرعرة) مقابل جبل دباغ، بلدية مجاز عمار على بعد بضعة كيلومترات غرب مدينة قالمة	-	23 أغسطس 1932	27 ديسمبر 1978
.73	علي عبد الله صالح	علي عبد الله صالح	بيت الأحمر بسنحان	-	21 مارس 1942	-
.74	الشريف حسين	حسين بن علي الهاشمي	إسطنبول	توفي سنة 1350 هـ - 1931م، فحمل إلى القدس ودفن فيها.	سنة 1270هـ	1854م
.75	سعود الفيصل بن عبد العزيز آل سعود	الأمير سعود بن فيصل بن عبد العزيز آل سعود	الطائف	-	2 يناير 1940	9 يوليو 2015

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.76	فيصل بن عبد العزيز آل سعود	فيصل بن عبد العزيز بن عبد الرحمن بن فيصل بن تركي آل سعود	-	-	13 ربيع الأول 1395	25 مارس 1975
.77	جابر الأحمد الصباح	جابر الأحمد الجابر الصباح	-	-	29 مايو 1926	15 يناير 2006
.78	موسى أبو مرزوق	موسى محمد أبو مرزوق	مخيم رفح	-	1951م	-
.79	وليد جنبلاط	وليد جنبلاط	المختارة بقضاء الشوف في لبنان	-	7 أغسطس 1949	-
.80	ميشال عون	ميشال عون	-	-	**17 فبراير 1932**	-
.81	حافظ الأسد	حافظ الأسد	مدينة القرداحة بمحافظة اللاذقية	-	6 تشرين الأول 1930	10 حزيران 2000
.82	عبد الكريم قاسم	عبد الكريم بن قاسم بن محمد بن بكر بن عثمان الفضلي الزبيدي	محلة المهدية على جانب الرصافة ببغداد	-	1914	1963

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.83	معمر القذافي	معمر محمد عبد السلام بن حُميد أبو منيار بن حُميد بن نايل القُحصي القذافي	)جهنم) بالقرب من (شعيب الكراعية) في وادي جارف بمنطقة سرت	-	7 يونيو 1942	20 أكتوبر 2011
.84	خليل الوزير	خليل إبراهيم محمود الوزير	الرملة بفلسطين	-	10 أكتوبر 1935	16 أبريل 1988
.85	سلام فياض	سلام خالد فياض خضر	دير الغصون قضاء طولكرم	-	1952	-
.86	طلال أبو غزالة	طلال أبو غزالة	يافا في فلسطين	-	22 أبريل 1938	-
.87	فاروق القدومي	فاروق رفيق الاسعد القدومي	جينصافوط إحدى قرى قلقيليه	-	ولد سنة 1931	-
.88	فيصل الحسيني	فيصل عبد القادر الحسيني	بغداد.	الكويت	(17 يوليو 1940	31 مايو 2001
.89	غسان الشكعة	غسان وليد الشكعة	نابلس	-	1943	-
.90	حنا ناصر	حنا ناصر	-	-	-	-
.91	هاني الحسن	هاني الحسن	حيفا	-	1939	6 يوليو 2012

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.92	حكم بلعاوي	حكم بلعاوي	-	-	-	stub -
.93	قيس عبد الكريم	قيس عبد الكريم	عراقي مقيم في رام الله	-	-	-
.94	ناصر القدوة	ناصر القدوة	بغزة	-	-	16 نيسان 1953
.95	ناصر اللحام	ناصر اللحام	مخيم الدهيشة للاجئين في مدينة بيت لحم الضفة الغربية	-	20 أبريل 1966	-
.96	غازي حمد	غازي حمد	يبنا قضاء فلسطين	-	1964	-
.97	نبيل عمرو	نبيل محمود عمرو	ولد بمدينة دورا التي تقع جنوب غربي الخليل	-	6 سبتمبر/أيلول 1947	-
.98	محمد الجوادي	محمد محمد الجوادي	مدينة فارسكور محافظة دمياط جمهورية مصر العربية	-	1958	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.99	عمر المختار	عُمر بن مختار بن عُمر المنفي الهلالي	البطنان ببرقة في الجبل الأخضر	-	20أغسطس1858	16سبتمبر 1931
.100	حسن البنا	حسن أحمد عبد الرحمن محمد البنا الساعاتي	المحمودية من أعمال محافظة البحيرة بدلتا النيل	-	14أكتوبر1906]	12فبراير 1949 م
.101	رياض الصلح	رياض الصلح	صيدا	-	1894	16يوليو 1951
.102	أنطون سعادة	أنطون سعادة	بلدة الشوير في جبل لبنان	-	1مارس 1904	8يوليو 1949
.103	زايد ال نهيان	زايد بن سلطان آل نهيان	أبوظبي بقصر الحصن	-	6)مايو1918	2نوفمبر 2004
.104	محمد أبو حامد	محمد أبو حامد شاهين	-	-	14 مارس1973	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.105	الخديوي اسماعيل	إسماعيل بن إبراهيم باشا بن محمد علي باشا	المسافر خانه بالجمالية	-	31ديسمبر 1830	2مارس1895
.106	محمد علي باشا	محمد علي باشا المسعود بن إبراهيم آغا القوللي	قولة التابعة لمحافظة مقدونيا	-	-	-
.107	فؤاد الأول	فؤاد بن إسماعيل بن إبراهيم باشا بن محمد علي باشا.	بقصر والده الخديوي إسماعيل بالجيزة،	بقصر القبة	26)مارس1868	28أبريل1936
.108	محمود سامي البارودي	محمود سامي بن حسن حسين بن عبد الله البارودي المصري	القاهرة	-	6أكتوبر1838	12 ديسمبر1904
- .109	أحمد فؤاد الثاني	-	القاهرة	-	16يناير1952	-

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.110	طلال بن عبد الله بن حسين	طلال بن عبد الله بن حسين الهاشمي	-	-	26فبراير 1909	7 يوليو 1972

Appendix 2: **Aljazeera Articles Dataset**

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.1	محمد نجيب	محمد نجيب يوسف	ساقية أبو العلا بالخرطوم	-	20فبراير 1901	28أغسطس 1984
.2	الملك فاروق	-	القاهرة	ايطاليا	11فبراير 1920	18 مارس 1965
.3	أنور السادات	محمد أنور السادات	ميت أبو الكوم التابعة لمحافظة المنوفية	-	25ديسمبر 1918	6 أكتوبر 1981
.4	جمال عبد الناصر	-	حي باكوس بالإسكندرية	-	15يناير 1918	28سبتمبر 1970

#	Name	Full Name	Birth Location	Death Location	Birth Date	Death Date
.5	الحسين بن طلال	-	عمان	-	14نوفمبر 1935	7فبراير 1999
.6	خليفة بن زايد آل نهيان	-	قلعة المويجعي بمدينة العين في إمارة أبو ظبي	-	1948	-
.7	زايد بن سلطان ال نهيان	-	أبوظبي	-	1918	2نوفمبر/تشرين الثاني 2004
.8	قابوس بن سعيد	-	مدينة صلالة بمحافظة ظفار	-	18نوفمبر 1940	-
.9	بسام الصالحي	بسام أحمد عمر الصالحي	رام الله	-	16يناير 1960	-
.10	فؤاد معصوم	محمد فؤاد معصوم هورامي	أربيل كردستان العراق	-	1يناير 1938	-