Natural Language Processing and Categorizing Arabic Articles (MVP)

Abeer Almdani , Sarah Alrashidi

Introduction

The goal of this project is to use Natural Language Processing and Unsupervised Learning models to categorize them into sections. After refining the approach. In this project, we will analyze the articles and categorize them into groups such as political, economic, medical, technological, scientific, sports and geographic articles.

Baseline

- **Latent Semantic Analysis Model (LSA)**
  In baseline we choose 1050 articles randomly to apply LSA, after vectorizing the data there are 16,575 word, in figure below The 8 topics, some of them are repeated.

```
Topic  1
الله , صلي, سلم , تعالي, قال, رسول , النبي, بن, رضي, انه

Topic  2
الشعر, يمكن, خلال, بشكل, الجسم , الوزن, الحمل, العديد, يجب, يتم

Topic  3
الشعر, لشعر, صحه , فروه , الراس, الطبيعيه , زيت, البيض, استخدام , الزيوت

Topic  4
الحمل, الوزن, خلال, الولاده , المراه , سلم , يمكن, الجسم , صلي, زياده

Topic  5
السلام , الصلاه , تعالي, الانبياء , القران, قال, النبي, الكريم , السنه , النبويه

Topic  6
الحمل, الشعر, الوزن, خلال, مدينه , الولاده , المراه , الصلاه , اسابيع , بدايه

Topic  7
الصلاه , صلي, سلم , السلام , النبي, خمسا , الدم , وسيدنا , ركعه , صلاه

Topic  8
الدم , الحمل, البيضاء , الشعر, خلايا , مدينه , انها , كريات, الجسم , الوزن
```

Figure 1: the 8 categories of articles

EDA phase

- normalize arabic گ → ك
- remove arabic numbers (٠،١،٢،….)
- remove diacritics (ً ٌ ٍ َ ُ ِ ّ ْ)
- remove punctuations (...،؟)
- remove repeating char
- remove arabic stopword
- remove non-arabic char
- remove unnecessary spaces

Models

- **Latent Semantic Analysis Model (LSA)**
  we used this model to categorize articles to many categories this model creates 4 categories that are beauty, religion, medicine and health and geography

- **Non-Negative Matrix Factorization (NMF)**
  we used this model to categorize articles to many categories this model creates 5 categories that are beauty, religion, medicine and health, geography and woman and child