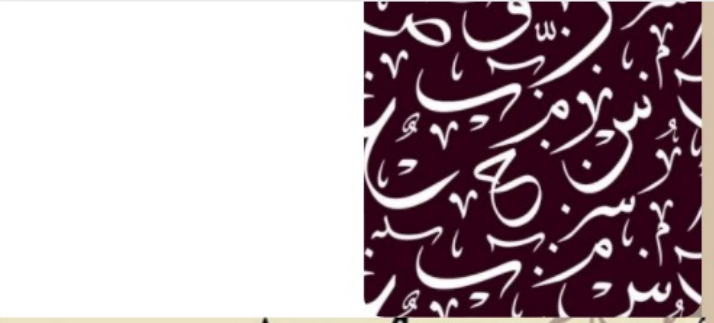

Natural Language Processing and Categorizing Arabic Articles

by: Abeer Almdani , Sarah Alrashidi

Outlines

1. **Introduction and business problem**
2. **Data**
    - EDA
    - Text preprocessing
3. **Algorithm**
    - Topic Modeling
    - Recommendation-systems
4. **Tools**
5. **Future work**

Introduction and Business problem

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

DATA

Data source Kaggle
https://www.kaggle.com/oussamaseffai/arabic-article-headline-generation

- 1050 rows
- 3 columns (title,introduction,index)

- Max words number is 1972 words
- Min words number is 116 words
- Max token length is 397 tokens
- Min token length is 19 tokens

DATA

Most Used Words

طريق	18.611316
علاج	15.534055
جسم	15.425522
بشر	15.257080
ماء	15.015648
زيت	13.401183
عام	13.044927
اعراض	12.822835
عمل	12.047188

bigrams n2

```
[('زيت زيت', 28)
('درج حرار', 27)
('احجار كريم', 24)
```

trigrams n3

```
[('زيت جوز هند', 12),
('ممارس تمار رياض', 11),
('كريا الدم بيضاء', 11),
```

DATA: Exploratory data analysis (EDA)

Remove NULLS and Duplicate values

Merge Title and Introduction columns

Remove Unnecessary columns

Removed short decomments

DATA: Text preprocessing

Remove Non-Arabic Char

Remove extra spaces

Remove Arabic Diacritics

Remove Arabic Numbers
٠،١،٢،٣،٤،٥،٦،٧،٨،٩

Normalize Arabic Letters
ك → گ

Remove Arabic Punctuations
؟ ، . / !

Remove Arabic Stop Word
(350 word)

Remove Repeating Char
وو ---> و
يي ---> ي

Algorithm: Topic Modelling

Embedding

- CountVectorizer
- TF-IDF

Modelling

- LatentDirichletAllocation(LDA)
- LatentSemanticAnalysis(LSA)
- Non-NegativeMatrixFactorization(NMF)
- CorEx
- Clustring and (Elbow curve and TSNE)

NMF | Count Vectorizer

Topic 0
عام , عمل , عالم , شخص , حيا , تم , تاريخ , فتر , بدء , عمر , يوم , خاص

Topic 1
مدين , دول , تقع , جنوب , تعتبر , مساح , بحر , شرق , عرب , جزير , سكان , غرب

Topic 2
جسم , غذاء , تناول , انس , الصح , طبيع , انواع , حرار , يعتبر , لذل , مواد , لان

Topic 3
اصاب , حال , علاج , اعراض , مرض , مصاب , شخص , عدوي , فيروس , جهاز , الدم , تظهر

Topic 4
طريق , ماء , يتم , استخدام , بشر , شعر , لمد , مكون , زيت , وصف , غسل , حصول

Results

We can notice that topics are:

Topic 0: تاريخ
Topic 1: جغرافيا
Topic 2: الغذاء والطعام
Topic 3: الطب
Topic 4: الصحة والجمال

Recommendation system

- **Document Similarity with Count vectorizer**

مقالك ينتمي إلى المقالات الجغرافية

********** صيد الأسماك في مصر **********

مصر من البلدان العربية الغنية بمصادرها السمكية سواء التي تعيش في المياه العذبة أو بالمياه المالحة ، و قد تم إنشاء الكثير من المشاريع السمكية بسبب توفر المياه على طول مجرى نهر النيل و قناة السويس ، حيث تتم تربية أنواع محددة من الأس ماك بقصد تصديرها إلى الخارج .

Recommendation system

1

2

3

********** موقع طروادة الأثري و الجغرافي **********

أجريت الحفريات الأولى للمدينة سنة 1870م ، و تعتبر بقاياها من أهم و أبرز مظاهر الاتصال بين حضارتي الأناضول ، و البحر الأبيض المتوسط ، و تقع طروادة على تلة تسمى هيسارليك (بالإنجليزية : Hisarlık) ، و تطل هذه التلة على السهل الممتد ع لى طول ساحل بحر إيجة التركي ، و على بعد 4 . 8 كم من المدخل الجنوبي إلى مضيق الدردنيل ، و قد كشفت الحفريات عن كونه ا منطقة كانت قد سكنت منذ 8000 عام ؛ و بسبب موقعها ، فقد عملت كجسر ثقافي يربط بين منطقة تروا س ، و البلقان ، و الأنا ضول ، و بحر إيجة ، و البحر الأسود ؛ و ذلك بسبب الهجرة و التنقل . يعتبر موقع طروادة الأثري ذا أهمية بالغة ؛ لفهم تطو ر الحضارة الأوروبية ؛ حيث إنه شاهد على تتابع الحضارات التي احتلت المنطقة لأكثر من 4000 سنة ، كما أن العديد من الآثار اليونانية ، و الرومانية ، تعكس خصائص المستوطنات التي أقامت فيها ، إضافة إلى أن موقعها يعد ذا أهمية ثقافية كبيرة ؛ و ذلك بسبب تأثيره في الأعمال الأدبية ، مثل : ملحمة إلياذة للمؤلف هوميروس (بالإنجليزية : Homer's Illiad) ، و الإنياد ة لفرجيل (بالإنجليزية : Virgil's Aeneid) . و تعتبر طروادة نقطة استراتيجية للمدخل الجنوبي للمضيق ، الذي يربط البح ر الأسود مع بحر إيجة ، عبر بحر مرمرة ، كما أنها مصدر لطريق بري يمتد شمالا على طول ساحل الأناضول الغربي ، و عبر أضيق نقطة من الدردنيل إلى الشاطئ الأوروبي ، و قد ظل الموقع الدقيق لهذه المدينة غير محدد حتى العصر الحديث ؛ حيث بدأت الح فريات لاكتشاف المدينة منذ عام 1870 ، و حتى عام 2005 م ، من خلال العديد من الحملات كان آخرها بقيادة عالم الآثار في جام عة توبنغن ، مانفريد كورفمان (بالإنجليزية : Manfred Korfmann) .

********** مميزات جزيرة المرجان **********

تحتوي هذه الجزيرة على مساكن مميزة تعرف بباب البحر ، و هي في معظمها عبارة عن وحدات فاخرة للسكن ، إضافة إلى فنادق عالمية كفندق ريكسوس التركي ، و الذي يضم ستمئة و خمسين غرفة ، و هو يعتبر من المنتجعات الضخمة ذات النجوم الخمسة ، و الذي بني على مبان ثلاثة ذات شكل هرمي ، و فيها قاعات خاصة بالمؤتمرات ، إضافة إلى المسرح ، و يوجد أيضا فندق هيلتون د بل تريه ، وكراون بلازا بفرعه الوحيد في دولة الإمارات على أرض هذه الجزيرة . يأتي في مقدمة هذه الفنادق منتجع ملدريد ، الذي تم منحه لأول مرة في العالم من قبل فريق ملدريد ، و يأتي على مساحة تقدر بأربعمئة ألف متر مربع ، و يتمتع بإطلا لة ساحرة ، و كذلك الأمر بالنسبة لفندق سبا جزيرة المرجان ، حيث يوجد فيه حوالي ثلاثمئة غرفة و جناح ، و قد صمم بطراز أ ندلسي ، و يتبع له كورنيش يبلغ طوله حوالي كيلومترا . تحوي أيضا جزيرة المرجان على عدد من المطاعم العالمية ، و تتوزع على شاطئ الجزيرة و كورنيشها المجهز بعدة مسارات ، بعضها مخصص لهواة رياضة المشي و حتى هواة المشي ، و بعضها الآخر لهو اة ركوب الدراجات الهوائية ، إضافة إلى العديد من المراكز الترفيهية في الجزيرة . الجدير بذكره أنه حسب تقرير لهيئة ت نمية السياحة في دولة الإمارات السنوي ، قد بلغت عائدات الدولة من إمارة رأس الخيمة في عام ألفين و أربعة عشر للميلاد م ا يزيد عن مليار درهم إماراتي ، و لعل لجزيرة المرجان اليد الطولى في الدعم الاقتصادي و الاستثماري في البلاد .

********** أكبر مدينة في العراق **********

تعد بغداد عاصمة دولة العراق و أكبر مدينة فيها و ثاني أكبر مدينة على مستوى الوطن العربي و غرب قارة اسيا ، حيث يبل غ عدد سكانها حوالي 7,180,889 نسمة ، تقع بغداد على طول نهر دجلة ، و بسبب موقعها المميز ازدهرت و تطورت المدينة بشكل سريع ، كما ساعدها موقعها من السيطرة على طرق التجارة الواقعة على طول نهر دجلة ، و يقسم نهر دجلة المدينة الى قسمين هما : الرصافة الواقعة في الجهة الشرقية ، و الكرخ الواقعة في الجزء الغربي ، و تعتبر مدينة بغداد مركزا للثقافة العر بية ، و تضم المدينة مجموعة من المعالم التاريخية و الأثرية السياحية أهمها : نصب الشهيد ، حديقة الحيوانات ، المتحف ا لوطني .

FUTURE WORK:

More Recommendation systems

Remove More Arabic Stop Word

Optimizing Arabic Python Libraries

Tools:

- **Software Platform:**
  - JupyterNotebook
- **Programming Language:**
  - Python
- **Python Libraries:**
  - Sklearn
  - Nltk
  - pyarabic.araby
- **Data manipulation libraries**
  - Pandas
  - Numpy
- **Visualization**
  - Matplotlib
  - Seaborn
  - Canva

THANKS!