Natural Language Processing and Categorizing Arabic Articles

by: Abeer Almdani , Sarah Alrashidi

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

Abstract:

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.

Design:

In this project, we will talk about articles in Arabic and categorize them into sections according to each category and the place to which this article belongs: history, geography, medicine, health and beauty and others.

Data:

data provided by kaggle:

- Max words number is 1972 words
- Min words number is 116 words
- Max token length is 397 tokens
- Min token length is 19 tokens
- Number of 1050 rows
- Number of columns 3
    - title
    - Introduction
    - index

Exploratory data analysis (EDA)

- Remove NULLS and Duplicate values
- Merge Title and Introduction columns
- Remove Unnecessary columns
- Add Artical column

Text preprocessing:

- Remove Non-Arabic Char
- Remove extra spaces
- Remove Arabic Diacritics
- Remove Arabic Numbers
- Normalize Arabic Letters گ → ك
- Remove Arabic Punctuations ؟ ، . / !
- Remove Arabic Stop Word (350word)
- Remove Repeating Char ي <--- يي , و <--- وو

Algorithm:

- **Embedding**
  - CountVectorizer
  - TF-IDF
- **Topic Modelling**
  - LatentDirichletAllocation(LDA)
  - LatentSemanticAnalysis(LSA)
  - Non-NegativeMatrixFactorization(NMF)
  - CorEx
  - Clustring and (Elbow curve and TSNE)

Best model is NMF with Count Vectorizer because it generates manful topics, but other models give unrelated terms for each topic

- Most common words in topic 1
  عام , عمل , عالم , شخص , حيا , تم , تاريخ , فتر , بدء , عمر , يوم , خاص
- Most common words in topic 2
  مدين , دول , تقع , جنوب , تعتبر , مساح , بحر , شرق , عرب , جزير , سكان , غرب
- Most common words in topic 3
  جسم , غذاء , تناول , انس , الصح , طبيع , انواع , حرار , يع , تبر , لذل , مواد , لان
- Most common words in topic 4
  اصاب , حال , علاج , اعراض , مرض , مصاب , شخص , عدوي , فيروس , جهاز , الدم , تظهر
- Most common words in topic 5
  طريق , ماء , يتم , استخدام , بشر , شعر , لمد , مكون , زيت , وصف , غسل , حصول

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
  Below, the article that we will predict similar articles to:

صيد الأسماك في مصر

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

- Similar articles:
  - First article:

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 يبلغ 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 مر كزا للثقافة العربية ، و تضم المدينة مجموعة من المعالم التاريخية و الأثرية السياحية أهمها : نصب الشهيد ، حديقة الحيوا نات ، المتحف الوطني.

  - Second article:

الرياض عاصمة السعودية

تعد مدينة الرياض (بالإنجليزية (Riyadh : العاصمة الرسمية للمملكة العربية السعودية و أكبر مدينة فيها ، و تعتبر من ا لمدن العربية الكبرى من حيث المساحة الجغرافية . تقع في المنطقة الوسطى بالقرب من المنطقة الشرقية على هضبة في ا لقسم الشرقي من هضبة نجد في منتصف شبه الجزيرة العربية، و يصل ارتفاعها فوق سطح البحر إلى 600م ، و تبعد الر ياض عن مدينة جازان مسافة 1245كم من الجهة الشمالية الشرقية ، و عن مدينة الدمام بحوالي 389كم ، و تبعد عن مكة ا لمكرمة مسافة 880كم من الجهة الشرقية ، و الآتي معلومات عن مدينة الرياض:

  - Third article:

موقع دولة العراق

تقع العراق في الجزء الجنوبي الغربي من قارة اسيا، وتشترك العراق في حدودها الشمالية مع دولة تركيا ، و يحدها من ال جهة الشرقية دولة إيران ، و من الجهة الجنوبية الكويت و المملكة العربية السعودية ، و سوريا و الأردن و المكلة العربية ا لسعودية من الجهة الغربية ، و تبلغ مساحة البلاد حوالي 438,317 كيلومترا مربعا ، و يصل عدد سكانها الى ما يقارب 3 2,585,000 نسمة حسب إحصاءات 2014م ، و تعتبر مدينة بغداد عاصمتها الرسمية ، و يعد مناخها جافا صحراويا في أ غلب مناطق الدولة.

Tools

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

Future Work

- More Recommendation systems
- Optimizing Arabic Python Libraries
- Remove More Arabic Stop Word