Assignment 1

12/10/2017 - Due 30/10/2017 CMP 614 - Text Mining

Clustering Turkish Thesis

In this assignment you are expected to implement a document-term vector space, apply weighting functions and use these for text clustering. We have two important goals, first you will practice and probably realize what we mean by vector space models by implementing it. Second, you will work with k-means algorithm and get used to considering documents as points and discover science domains in Turkish thesis dataset.

Dataset

YÖK serves all Turkish thesis from a website. I will provide this data to you in a password-protected zip file from the course web-site and email you the password.

`https://tez.yok.gov.tr/UlusalTezMerkezi/`

The data is coded as a JSON collection so that you can access this data easily with a JSON parser or with regular expressions.


```
"_id": {
"$oid": "5993737a0fb6335388779f3f"
},
"meta": "444639\u0009\nTopic model based recommendation systems for retailers / Satıcılar için konu
 modelleme yöntemine dayalı öneri sistemi\nYazar:RIMA TÜRKER
 \nDanışman: YRD. DOÇ. DR. GÖNENÇ ERCAN\nYer Bilgisi: Hacettepe Üniversitesi / Fen Bilimleri Enstitüsü /
 Bilgisayar Mühendisliği Anabilim Dalı\
 nKonu:Bilgisayar Mühendisliği Bilimleri-Bilgisayar ve Kontrol = Computer Engineering and Computer
 Science and Control ; Bilim ve Teknoloji = Science and Technology
 \nDizin:\u0009 Onaylandı\nYüksek Lisans\nIngilizce\n2016\n61 s.",
"tr": "Günümüzde, satıcıların daha fazla müşteri kazanmak, var olan müşterilerinin sadakatını
sürdürmek ve artırabilmek için iyi bir stratejiye ihtiyaçları vardır. Bu görevi gerçekleştimek
için en iyi yollardan birisi müşteri profiline uygun ve müşterinin ihtiyacını karşılayabilecek
yeni ürünleri (daha önce satışı bulunmayan) müşterilere sunmaktır. Bu tezde, satıcılara yardımcı
olabilmek için yeni ürün önerme yeteneğine sahip bir öneri sistemi geliştirilmiştir. Klasik öneri
 sistemleri müşteriye ürün önermek için geliştirilmiştir, ancak bu çalışmada geliştirilen sistemin
  en büyük farklarından birisi müşteriye değil satıcıya yönelik yeni ürünleri önerebilen bir öneri
   sistemi geliştirmektir. Söz konusu sistemi geliştirebilmek için \"Olasalıksal Örtük Anlam
   Analizi\" metodu genişletilmiştir. Genişletilen metodun asıl amacı müşterileri alışveriş verilerini
   kullanarak satıcıların satma olasılığı yüksek olan yeni ürünleri tespit etmektir. Bu amaç
   doğrultusunda üç temel veri kaynağı dikkate alınmıştır; müşteri, müşterinin ziyaret ettiği veya
   alışveriş yaptığı satıcı ve müşterinin aldığı ürün. Bahsedilen veri kaynakları kullanılarak ilgili
    değişken olasılıkları hesaplayabilmek için olasılıksal model geliştirilmiştir. Bu modelin
    doğrulanabilmesi için gerçek müşteri veri setlerini kullanarak deneyler yapılmıştır. Ayrıca
     deney sonuçlarının karşılaştırılması ve daha uygun değerlendirme yapılabilmesi için
     \"Işbirliğine Dayalı Filtreleme\" methodu bazal algoritma olarak kullanılmıştır. Aynı deney
     koşulları sağlanarak iki algoritma aynı verilerle test edilmiş ve sonuçlar karşılaştırılmıştır.
      Test neticelerine göre bu tez doğrultusunda tasarlanan modelden daha doğru ve gerçeğe yakın
       sonuçlar alındığı gözlemlenmiştir.",
"tezNo": 444639
}
```


Each thesis consists of 3 fields and you will be using the text in the field "tr". Using the Turk-

ish abstract of the thesis you will try to find how many different scientific domains are there in the dataset. The field meta stores some metadata from the site, we are trying to find clusters corresponding to general scientific domains as in the metadata "Yer Bilgisi".

For this task you will use vector spaces with tf-idf weights. You may use stemming and stop-word removal, but this is not a must. You will implement k-means++ algorithm. You will perform clustering with different k values and try to find an appropriate k value using the elbow method. After finding an appropriate k value you will calculate the clustering purity of 4 randomly picked clusters. Clustering purity requires ground truth category information for the documents, you can use the last portion of the Yer Bilgisi field for this purpose. For the example above the category of the document will be "Bilgisayar Mühendisliği Anabilim Dalı". You can manually check the labels if the number of documents is small.

You have to submit:

- Your source code for the assignment
- A report containing
  - If you have done something different from white-space tokenization explanation of you procedure
  - Any modifications you did on the k-means algorithm like k-medoid etc.
  - Your plot for k-values and squared errors.
  - Result of clustering purity for the selected clusters and dominant category for each cluster.
- A Readme file for instructions on compiling and running your code.

Submission

Your assignment is due for 30/10 Monday 23:59. You have 3 late day submissions, with a penalty of 10% per day. Please try to include short and clear instructions for the code, a small readme file. You can submit your assignment to the web-page given below (you can update your submission as well), zip all the documents and upload.

`https://script.google.com/macros/s/AKfycbyP-qNzlBtVYTFlrysdXkhnMmrYikJVBRE5yyEOJMT4QqT03Snh/exec`

Enjoy...