CS5560 Knowledge Discovery and Management

Problem Set 5

February, 18th 2020

Name:

Class ID:

1. **Information Retrieval (Text Mining) with TF-IDF**

Consider the following three short documents

Doc #1:

The researchers will focus on computational phenotyping and will produce disease prediction models from machine learning and statistical tools.

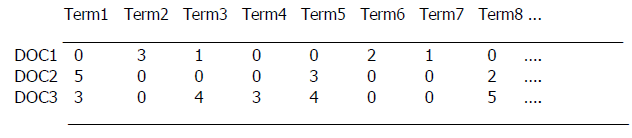
Doc #2:

The researchers will develop tools that use Bayesian statistical information to generate causal models from large and complex phenotyping datasets.

Doc #3:

The researchers will build a computational information engine that uses machine learning to combine gene function and gene interaction information from disparate genomic data sources.

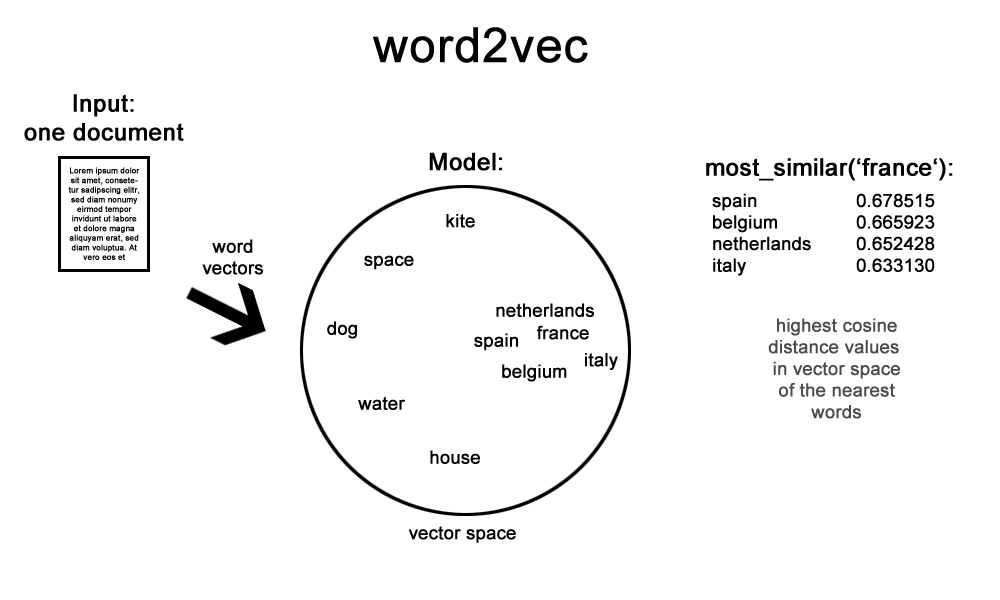
1. First remove stop words and punctuation; detect manually multi-word terms (using N-Gram or POS Tagging/Chunking); parse manually the documents and select the terms from the given 3 documents and created the dictionary (list of terms).
2. Create the document vectors by computing TF-IDF weights. Show how to compute the TF-IDF weights for terms. For each form of weighting list the document vectors in the following format:



1. **Word2Vec**

Word2Vec reference: <https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/>

Consider the following figure showing the Word2Vec model.



1. Describe the word2vec model
2. Describe How to extend this model for multiple documents. Also draw a similar diagram for the extended model.

Describe the differences of the following approaches

* Continuous Bag-of-Words model,
* Continuous Skip-gram model

For the sentence “morning fog, afternoon light rain,”

* Place the words on the skip-gram Word2Vec model below.
* Draw a CBOW model using the same words.

