Lab 4 - Answer suggestions

Topic Analysis

(pages 304-308, 325-342)

1. What is the difference between supervised and unsupervised learning? Discuss some benefits and issues for each approach in the context of topic analysis.

- In short, supervised learning requires labeled data, while unsupervised learning does not. The distinction is quite clear, although there are cases where an unsupervised approach can be used to *create* labels, called semi-supervised learning. For topic analysis, an unsupervised approach may be more applicable, as topics are rarely well-defined for all domains. A benefit of the unsupervised approach is that we can apply it to any corpus, whereas a supervised topic model will likely perform poorly on out-of-domain tasks. A clear benefit of supervised learning is that we can evaluate it with clearly defined metrics, such as precision, recall, and f-measure. For unsupervised, we can use clustering-related metrics (listed later).

2. You are given a corpus of 1 million documents and a vocabulary of 100,000 words. List some problems you may encounter when using TF-IDF vectorization for clustering this corpus, and explain how you would deal with them.

- Essentially we end up with a massive matrix and will run into memory issues for computing TF-IDF. The best way to deal with this is to reduce the dimensionality, e.g., by using singular value decomposition (SVD), and/or by more simple terms of removing stopwords and other frequency-based measures.

3. Metrics are essential when dealing with machine learning. However, regarding unsupervised clustering (e.g., of topics), we cannot use the typical precision, recall, and f-measure metrics. What are the alternatives for this task?

- Homogeneity, completeness, v-measure, ...

Topic Modeling

(pages 349-356, 371-377)

Given the five sentences:

> "Macrosoft announces a new Something Pro laptop with a detachable keyboard."
>
> "Melon Tusk unveils plans for a new spacecraft that could take humans to Mars."
>
> "The top-grossing movie of the year Ramvel Retaliators."
>
> "Geeglo releases a new version of its Cyborg operating system."
>
> "Fletnix announces a new series from the creators of Thinger Strangs."

1. How would *you* (without programming) assign the listed sentences to separate topics? Explain your chain of thought.

- This is an open question!

2. Two algorithms for topic discovery are Latent Semantic Indexing (LSI) and Latent Dirichlet Allocation (LDA)

- What preprocessing steps should we consider before implementing these algorithms?
  - Stemming, lemmatization, stopword removal, TF-IDF
- Both algorithms require the user to specify the number of topic clusters. How can we *automatically* detect a reasonable number of topics? Tip: Look into metrics for unsupervised clustering.

  - The elbow method is a common approach to evaluate the suitable amount of clusters. There are others, though, such as the silhouette score. Read more here: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

3. Using the corpus from Lab 3: `amazon_appliances_reviews`, implement an LSI and LDA model using Gensim. Specify 5 topics and print out the top 10 words for each topic. The package `pyLDAvis` will help you to visualize the topics!

- Which algorithm do you think is more accurate? Why?
  - Open question for the student.
- How do the topics compare to your interpretations? In your opinion, are there more or less than 5 *actual* topics?
  - Open question for the student.

Named Entity Recognition

(pages 384-392, 403-415)

1. In Lab 3, you learned about noun phrases. Noun phrases, such as "The quick brown fox" or "Mount Everest", are often considered named entities. Give examples of named entity categories that are *not* noun phrases.

- Entities can occur in many forms:
  - ORGANIZATION - Georgia-Pacific Corp., WHO
  - PERSON - Eddy Bonte, President Obama
  - LOCATION - Murray River, Mount Everest
  - DATE - June, 2008-06-29
  - TIME - two fifty a m, 1:30 p.m.
  - MONEY - 175 million Canadian Dollars, GBP 10.40
  - PERCENT - twenty pct, 18.75 %
  - FACILITY - Washington Monument, Stonehenge
  - GPE - South East Asia, Midlothian

  Here, DATE/TIME/MONEY are examples that may occur as non-NP entities.

2. Disambiguating (or entity linking) named entities is a crucial task to applications of NER and considers the problem of assigning an identifier to each entity, i.e., *linking* relevant entities together. The disambiguation process often incorporates external knowledge (*knowledge bases*).

   Consider the sentences:

   > "I ate an apple in New York"
   >
   > "New York Times wrote an article about Apple"
   >
   > "New York is also known as the Big Apple"

   How would you tackle the task of distinguishing the entities found here? Give a rough step-by-step explanation, either in text or by pseudo-code — no implementation required.

   - Knowledge bases often include information that specifies the *type* of entity, e.g., a person. A decent NER system may also correctly label this. Regardless, you could identify "New york times" as having an organization number and "New York" as a location. You could also use a knowledge base to identify "Apple" as a company and "Big Apple" as a nickname for New York.