GENERAL ASSEMBLY

INTRO TO DATA SCIENCE
LECTURE 13: NATURAL LANGUAGE PROCESSING

Paul Burkard
08/06/2015

LAST TIME:

- DIMENSIONALITY REDUCTION
- FEATURE SELECTION: WRAPPERS, FILTERS, EMBEDDED
- FEATURE EXTRACTION: PCA, SVD

QUESTIONS?

I. NLP

HANDS-ON: LATENT SEMANTIC ANALYSIS

- What is NLP?
- What are some considerations in NLP?
- What is Latent Semantic Analysis?
  - How do we perform LSA?

I. NLP

Q: *What is **NLP**?*

A: *A field of computer science, artificial intelligence, and linguistics concerned with the interaction between computers and human languages.*

The goal is for computers to derive meaning from human natural language input.

There are some general considerations that NLP faces...

Q: *What is* ***tokenization****?*

A: *Tokenization is the process of breaking up streams up text into words, phrases, symbols, or other meaningful elements called* ***tokens***

These tokens become inputs for further ML processing and generally allow us to put text information into data vectors (a vector space, this "vectorizing" of the data is always the first step in any ML problem).

Q: *What is* ***stemming****?*

A: *Stemming is reducing words to that share the same root (verb forms, plurals, etc) to their root word for further processing.*

The idea is that the semantic information is captured by the root, and that retaining all the different forms just adds noise and complexity.

Q: *What is* ***TFIDF*** *weighting?*

A: Term-Frequency Inverse-Document-Frequency *assigns a weighting scheme that is proportional to the frequency of a token within a group of words and inversely proportional to the token frequency in the entire set of documents (corpus).*

We will apply TFIDF weighting to generate better vectors for our ML algorithms.

Q: *What is a* ***Bag of Words Model****?*

A: *The idea that the order of the words in a document don't matter much in terms of semantic or conceptual meaning of the document.*

This will allow us to ignore word order and just treat each possible token as a static coordinate in a vector space (with (possibly weighted) word frequencies as the coordinate values).

Q: *What is a* ***Cosine Similarity****?*

A: *A common metric for similarity used in vector spaces resulting from bag of words models on text analysis.*

It's the dot product of 2 vectors divided by the norms of the 2 vectors:

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Q: *What is a* ***Latent Semantic Analysis****?*

A: *LSA is a technique that combines an appropriate TFIDF weighting with a bag of words model to vectorize many documents of text data.*

The resulting ***Term-Document Matrix*** *is then reduced using an* ***SVD****.*

The output yields reduced Term and Document vector spaces which allow Term-Term, Term-Document, and Document-Document similarity comparisons via cosine similarity in the (drastically) reduced-dimensionality space.

HANDS-ON: LSA (LSI)