Day 4 Notes

Full text indexing

Bag of Words

- Adjacency representation of tokenized words and documents they exist in
- Allows you to represent the text as some form of distance
- **Pros**: simple, very commonly used
- **Cons**: No grammer, no order, want to preserve as much info as possible - indexes of words that exist, etc.

N-Gram

- Continguous sequence of n items from a given sequence.
- Eg. info retrieval is . . . .
- Bigram: ‘info_retrieval’, ‘retrieval_is’,etc.
- **Pros**: local dependency and order
- **Cons**: increases vocab size to store representation O(V^N)

Full Text indexing

- Index doc with all occuring words
- Preserves info and automatic, but requires O(V^N) space and has vocab gap

Zipf’s law

- freq of any word is inversely proportional to its rank
- long tail - (?) power law
- –add equation–
- head words take large portion of occurences, but are semantically meaningless
- tail words, take major portion of vocab but rarely occur
- rest is most representative

Automatic text indexing

1) Remove non-informative words. Remove 1s. Using a linkedlist is more effecient since each node is 1s, the 0s aren’t even stored
2) Remove rare words. Remove 0s. If matrix/2d array, this is more effecient because there are more 0s.
3) Use the middle section

Why? For saving space.

Stopwords

- Useless words for query/doc analysis.
- Not all words are informative
- Remove such words to reduce vocab size
- No universal definition, have to define your own
- You risk breakig the original meaning and structure of the text

Normalization

- Convert different forms of a word to normalized form in the vocab
- Solution: Rule based & Dictionary based
- **Rule based**: Delete periods,etc. all in lower case
- **Dictionary-based**: Construct equivalent class. Car->“vehicle, auto” , etc.

Stemming

- Reduce inflected o derived words to root form.

- Plurals, adverbs, etc.
- Ladies-> lady, referring-> refer
- Risk: You might lose context, meaning again

Search Engine Arch so far

1) Crawler
2) Doc analyzer
3) BagOfWord representation

–write out review here–

Modern engine?

- No stemming or stopword removal. Storage isn't that big of an issue
- More and more into NLP techniques