Lect 7: Stop word-Stemming -Lemmatization

Dr. Subrat Kumar Nayak
Associate Professor
Dept. of CSE, ITER, SOADU

Stop Words

- Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called ***stop words.***
- The general strategy for determining a stop list is to sort the terms by ***collection frequency*** and then to take the most frequent terms.
- Often hand-filtered for their semantic content relative to the domain of the documents being indexed, as a ***stop list***, the members of which are then discarded during indexing.

a	an	and	are	as	at	be	by	for	from
has	he	in	is	it	its	of	on	that	the
to	was	were	will	with					

► **Figure 2.5** A stop list of 25 semantically non-selective words which are common in Reuters-RCV1.

- Using a stop list significantly reduces the number of postings that a system has to store.

Stop Words

- And a lot of the time not indexing stop words does little harm: keyword searches with terms like the and by **don't seem** very useful.
- However, **this is not true for phrase searches**.

Example:

- The phrase query "President **of the** United States", which contains two stop words, is more precise than President AND "United States".
- The meaning of flights **to** London is likely to be lost if the word **to** is stopped out.
- Some special query types are disproportionately affected. Some song titles and well known pieces of verse consist entirely of words that are commonly on stop lists.

Example: *To be or not to be, Let It Be, I don't want to be, . . .*

Stop Words

- With a stop list, you exclude from the dictionary entirely the commonest words. Intuition:
- They have little semantic content: *the, a, and, to, be*
- There are a lot of them: ~30% of postings for top 30 words

- But the trend is away from doing this:
- Good compression techniques (IIR 5) means the space for including stop words in a system is very small
- Good query optimization techniques (IIR 7) mean you pay little at query time for including stop words.
- You need them for:
  - Phrase queries: "King of Denmark"
  - Various song titles, etc.: "Let it be", "To be or not to be"

Lemmatization

- Reduce inflectional/variant forms to base form
- E.g.,
  - *am, are, is* $\rightarrow$ *be*
  - *car, cars, car's, cars'* $\rightarrow$ *car*
- *the boy's cars are different colors* $\rightarrow$ *the boy car be different color*
- Lemmatization implies doing “proper” reduction to dictionary headword form (the lemma)

Stemming

- Reduce terms to their “roots” before indexing
- “Stemming” suggests crude affix chopping
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.

for example compressed and compression are both accepted as equivalent to compress.

for exampl compress and compress ar both accept as equival to compress

Porter's algorithm

- Commonest algorithm for stemming English
  - Results suggest it's at least as good as other stemming options
- Conventions + 5 phases of reductions
  - phases applied sequentially
  - each phase consists of a set of commands
  - sample convention: *Of the rules in a compound command, select the one that applies to the longest suffix.*

Typical rules in Porter

- *sses* → *ss*
- *ies* → *i*
- *ational* → *ate*
- *tional* → *tion*

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

- Weight of word sensitive rules
- *(m>1) EMENT* → (m: length of rest word before ement)
  - *replacement* → *replac*
  - *cement* → *cement*

Other stemmers

- Other stemmers exist:
  - Lovins stemmer
    - http://www.comp.lancs.ac.uk/computing/research/stemming/general/lovins.htm
    - Single-pass, longest suffix removal (about 250 rules)
  - Paice/Husk stemmer
  - Snowball
- Full morphological analysis (lemmatization)
  - At most modest benefits for retrieval

Three stemmers: A comparison

- **Sample text**: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation
- **Porter stemmer**: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a picture of express that is more biolog transpar and access to interpret
- **Lovins stemmer**: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres
- **Paice stemmer**: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Language-specificity

- The above methods embody transformations that are
  - Language-specific, and often
  - Application-specific
- These are “plug-in” addenda to the indexing process
- Both open source and commercial plug-ins are available for handling these

Does stemming help?

- English: very mixed results. Helps recall for some queries but harms precision on others
  - E.g., operative (dentistry) ⇒ oper
  - operational(research)=> oper
  - operating(systems)=> oper
- Definitely useful for Spanish, German, Finnish, …
  - 30% performance gains for Finnish!