Domain Adaptation for Sentiment Classification

ErikAnthony Harté
Sentiment Analysis
Ling 575 Spring Quarter

Project Goal

- Examine domain adaptability from Amazon book reviews to IMDB movie reviews by characterizing linguistic style differences between the two domains, e.g.:
  - document length, sentence length, word frequency
  - distribution of POS, such as determiners and prepositions
  - use of subjectivity expressions
  - words with high information gain
  - cosine similarity

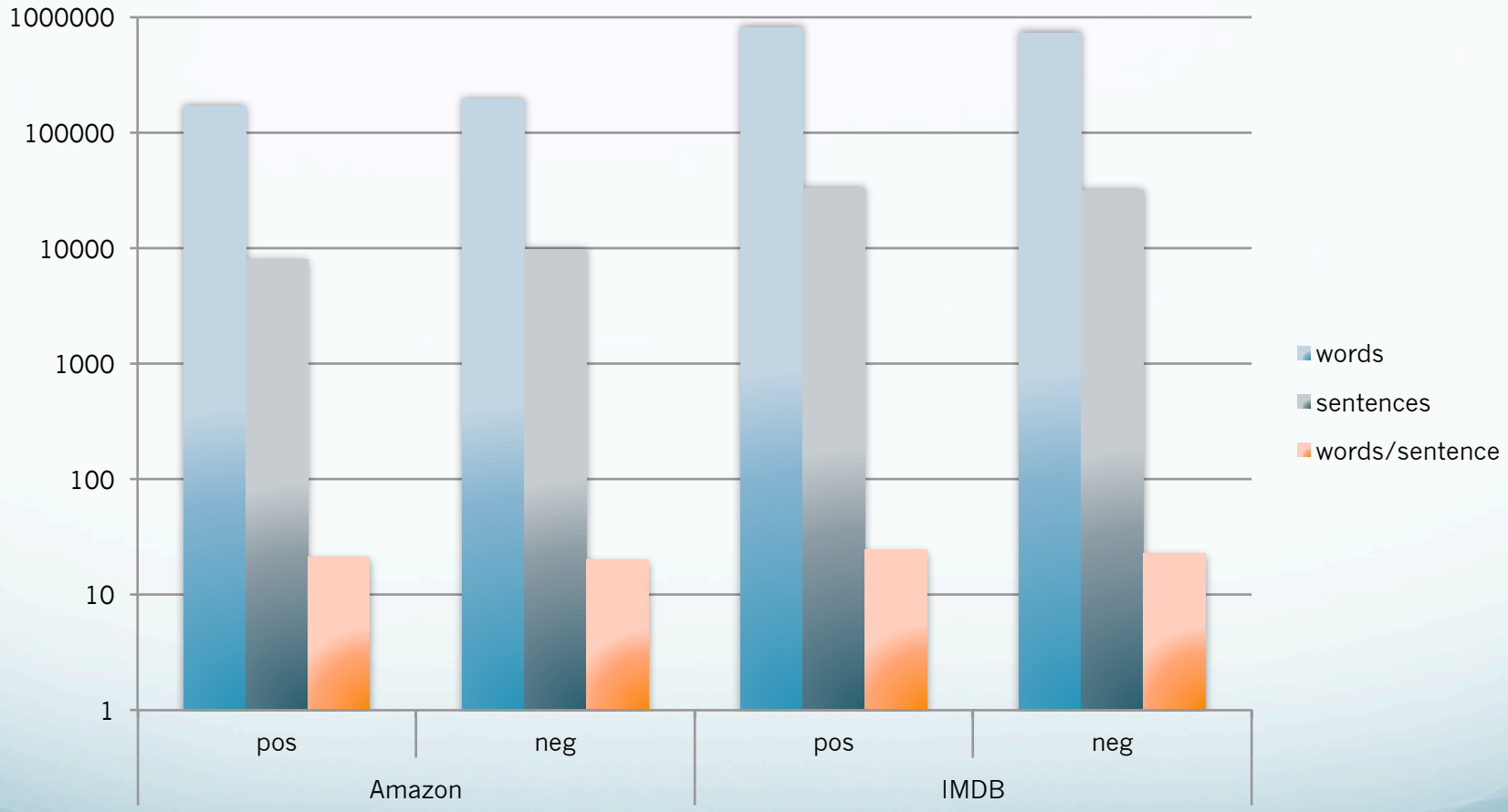
Dataset

- Amazon book reviews
  - 1K pos / 1K neg Amazon reviews. Mark Drezde & Richard Johansson
- IMDB movie reviews
  - 1K pos / 1K neg IMDB movie reviews. Bo Pang & Lillian Lee
- Subjectivity Lexicon
  - MPQA Subjectivity Lexicon. ~8K terms

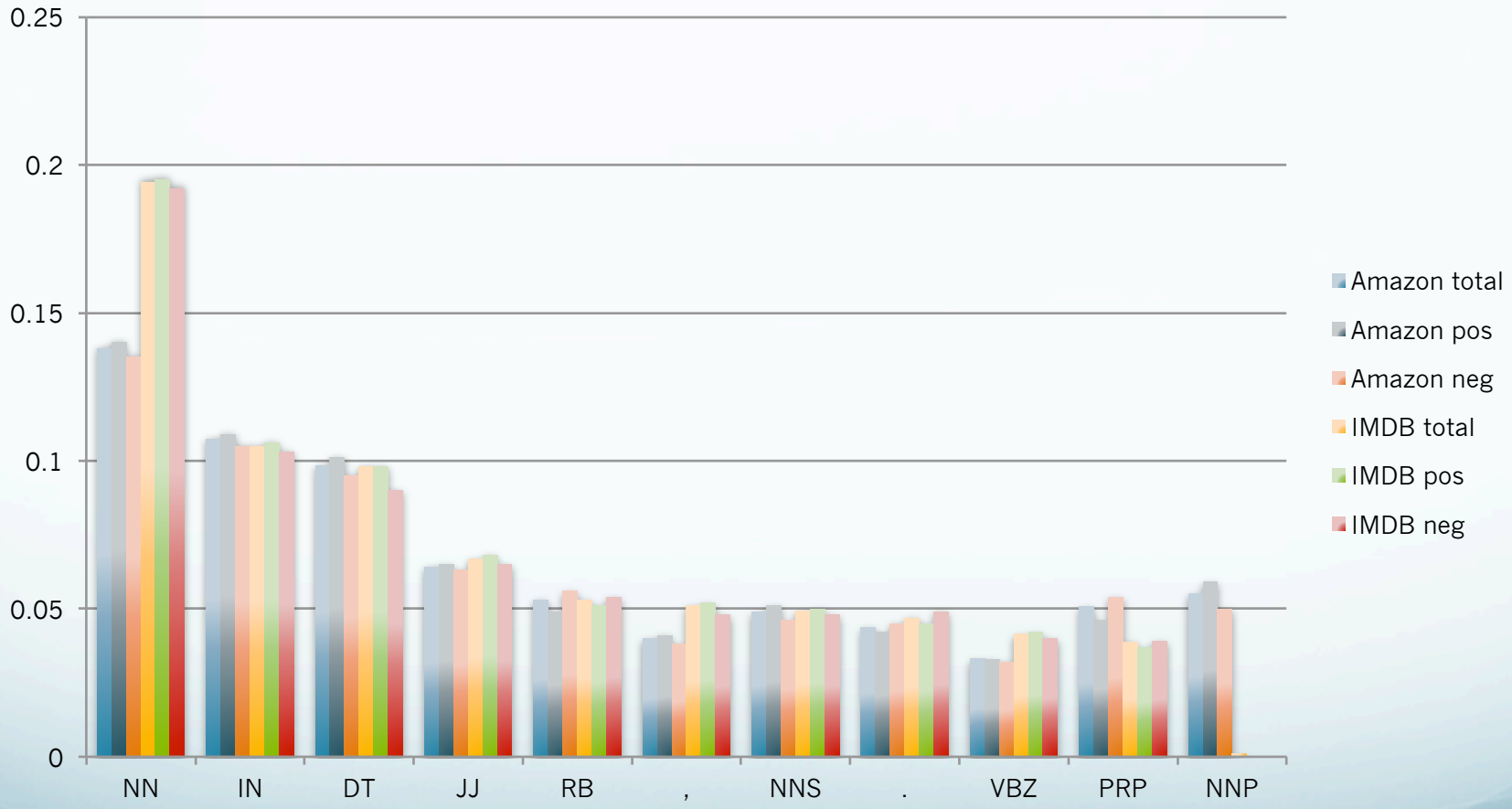
Tools

- MXTerminator
  - Sentence Boundary Detector. Adwait Ratnaparkhi
- NLTK Toolkit
  - Tokenizer
  - POS tagger
  - Chi-Square

Raw Counts


1000000
100000
10000
1000
100
10
1
pos
neg
pos
neg
Amazon
IMDB
words
sentences
words/sentence

POS Tagging

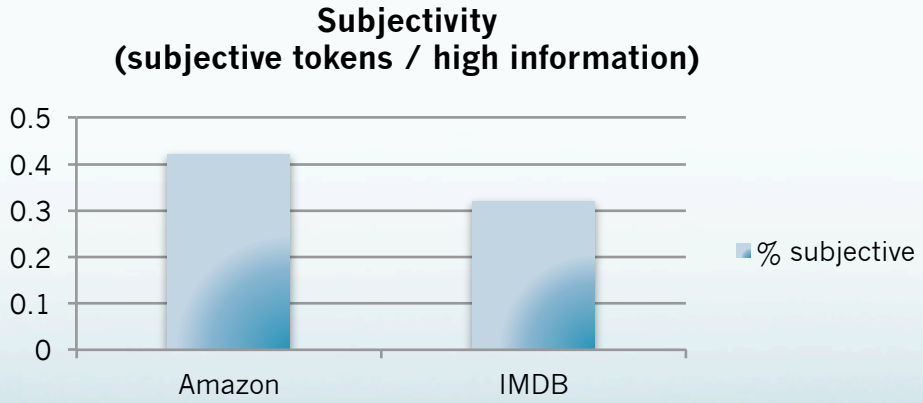

0.25
0.2
0.15
0.1
0.05
0
NN
IN
DT
JJ
RB
,
NNS
.
VBZ
PRP
NNP
Amazon total
Amazon pos
Amazon neg
IMDB total
IMDB pos
IMDB neg

Information Gain

- Amazon
  - **and**, heart, just, ..., money, ``, highly, not, \x1a, best, if, would, author, poorly, **there**, better, also, wonderful, too, recommend, **waste**, was, **war**, ?, **do**, **life**, finish, that, each, maizon, but, wrong, **worst**, excellent, **nothing**, "m", disappointed, pages, **great**, unfortunately, anything, **i**, **boring**, no, favorite, "**nt**", did, **bad**, of, or
- IMDB
  - **and**, **there**, is, **life**, terrible, wasted, as, nbsp, have, truman, **if**, !, plot, flynt, mulan, no, movie, *, ,', dull, ., **waste**, **war**, ?, **do**, ``, his, godzilla, mess, batman, seagal, ridiculous, **worst**, awful, why, **great**, lame, **i**, **boring**, script, jackie, "**nt**", this, **bad**, stupid, **nothing**, supposed, harry, the, shrek

Subjectivity

- Amazon
  - heart, just, best, would, poorly, better, wonderful, too, recommend, **waste**, **war**, **life**, wrong, **worst**, excellent, disappointed, **great**, unfortunately, **boring**, favorite, **bad**
- IMDB
  - **life**, terrible, plot, dull, **waste**, **war**, mess, ridiculous, **worst**, awful, **great**, lame, **boring**, **bad**, stupid supposed



Cosine Similarity

- ...still in progress ☺

Summary

- Documents in Movie domain 10x larger than in Book domain
- Book reviews have a substantially greater proportion of NNP. Movie reviews show greater proportion of NN.
- Unmodified information gain results not useful. Suspect filtering of named entities (e.g. ‘mulan’, ‘seagal’, ‘batman’, ‘godzilla’) might help
- Book reviews make greater use of subjectivity lexicon