Adminstrivia

• Announcement: I'm sick so this is going to be a very lecture
• Surprised so many of you submitted the optional homework 6 ...
Homework 6 review

• Read the review paper:
  
  **Zipf’s word frequency law in natural language: A critical review and future directions**, Piantadosi, S.


• Optional Homework 6
  
  • Using Python and our ptb WSJ corpus, compute and graph frequency against rank
  • Use pyplot
  • In your opinion, does our WSJ corpus obey Zipf's Law?
Homework 6 review

```python
import nltk
from nltk.corpus import ptb
from collections import Counter
from math import log, log10
import matplotlib.pyplot as plt
from sys import argv

n = int(argv[1]) if len(argv) == 2 else 100

excluded = set([-NONE-, -LRB-, -RRB-, SYM, :, ., ,, "", "]

for x in ptb.tagged_words(categories=['news']) if x[1] not in excluded

size = len(tokens)
```
Homework 6 review

```python
excluded = set(['-NONE-', '-LRB-', '-RRB-', 'SYM', ':', '.', ',', '', '``', '''])
tokens = [x[0] for x in ptb.tagged_words(categories=['news']) if x[1] not in excluded]
size = len(tokens)
c = Counter()
for token in tokens:
    c[token] += 1
mc = c.most_common()
print('corpus check: tokens:{}, types:{}, top5:{}'.format(size, len(mc), mc[:5]))
```
Homework 6 review

```python
ranks = [x for x in range(1, len(mc)+1)]
freq = [item[1]/size for item in mc]
plt.figure(1)
plt.xscale('log')
plt.yscale('log')
plt.grid(True)
plt.title('Log-log plot for freq vs rank for PTB WSJ for top-{} words'.format(topN))
plt.plot(ranks[:topN], freq[:topN])
```
Homework 6 review

```python
27 lrank = [log(x) for x in range(1, len(mc)+1)]
28 lfreq = [log(item[1]/size) for item in mc]
29 plt.figure(2)
30 plt.xscale('linear')
31 plt.yscale('linear')
32 plt.grid(True)
33 plt.title('Plot of log(freq) vs log(rank) for PTB WSJ for top-{} words'.format(topN))
34 plt.plot(lranks[:topN], lfreq[:topN])
35 plt.show()
```
Homework 6 review
Homework 6 review
Homework 6 review
Homework 6 review
Homework 6 review