Computing lexical diversity of tweets

Lexical Diversity

- What is it?
- Calculating simple frequencies and can be applied to unstructured text is a metric called *lexical diversity*.
- Mathematics?
- Number of *unique* tokens in the text divided by the *total* number of tokens in the text.
-

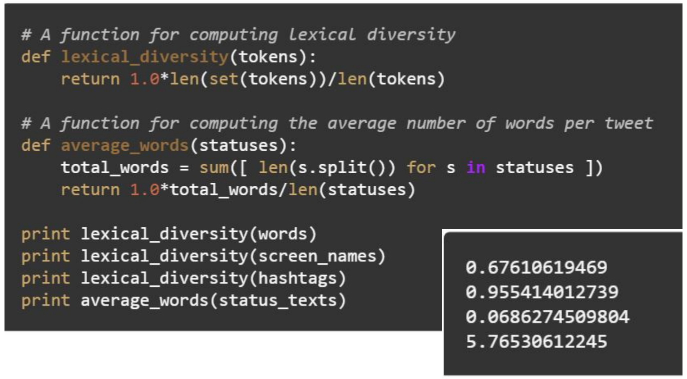
Lexical Diversity – II

- **Interesting concept in the area of interpersonal communications. Why?**
- It provides a quantitative measure for the diversity of an individual's or group's vocabulary.
- **An example : "And Stuff"**
- **Lexical diversity can be worth considering as a primitive statistic for answering a number of questions. How?**
- How broad or narrow the subject matter is that an individual or group discusses

Lexical Diversity – III

- **Breaking down the analysis to specific time periods could yield additional insight.**
- **Comparing different groups or individuals**
- **Lexical Diversity of Coca Cola and Pepsi**

An Example



```
# A function for computing lexical diversity
def lexical_diversity(tokens):
    return 1.0*len(set(tokens))/len(tokens)

# A function for computing the average number of words per tweet
def average_words(statuses):
    total_words = sum([ len(s.split()) for s in statuses ])
    return 1.0*total_words/len(statuses)

print lexical_diversity(words)
print lexical_diversity(screen_names)
print lexical_diversity(hashtags)
print average_words(status_texts)
```

```
0.67610619469
0.955414012739
0.0686274509804
5.76530612245
```

Understanding the Example

- Obs 1: 0.67: One in 3 words is a unique word.
- Obs 2: 0.97: About 19 out of 20 screen names mentioned are unique.
- Obs 3: 0.068: Diversity of hashtags very low.
- Obs 4: The average number of words per tweet is very low at a value of just under 6, which makes sense given the nature of the hashtag, which is designed to solicit short responses consisting of just a few words.

The final slide

- Given an average number of words per tweet as low as 6, it's unlikely that users applied any abbreviations to stay within the 140 characters, so the amount of noise for the data should be remarkably low, and additional frequency analysis may reveal some fascinating things.