Lecture 6: Preparing Spam Data

อ.ปรัชญ์ ปิยะวงศ์วิศาล

Today's Topics

Lab Exercise

- Use TextBlob to process spam data

From last lab…

Say, we want to split this message:

“Free entry in 2 a wkly comp to win FA Cup”

into a list of words (tokens):

[Free, entry, in, 2, a, wkly, comp, to, win, FA, Cup]

How do we do it the easy way?

Using split() to tokenize words – Any issues?

1. Create split function:

```
def split_words(message):
    return msg.split(' ')                    #  problem with this method?
```

2. Apply **split_words()** to messages:

```
messages.message.head().apply(split_words)
```

NLP questions?

- Do capital letters carry information?
- Does distinguishing inflected form ("goes" vs. "go") carry information?
- Do interjections, determiners carry information?

Installing TextBlob

1. Open Anaconda prompt
2. Type

```
pip install –U textblob

Python –m textblob.download_corpora
```

Simple NLP with TextBlob

Part-of-Speech Tags

```
TextBlob(“Hello, how is it going?”).tags
```

* http://www.ling.upenn.edu/courses/Fall_2007/ling001/penn_treebank_pos.html

Convert to lemmas (base form)

```
words = TextBlob(“Hello, how is it going?”).words

lemmas = [word.lemma for word in words]
```

More on what TextBlob can do

TextBlob Quick Start will walk you through basic commands for text processing

>> https://textblob.readthedocs.io/en/dev/quickstart.html <<

Split messages with TextBlob

1. Modify the split function to use TextBlob:

```
from textblob import TextBlob


def split_words(message):
    return TextBlob(message).words
```