A Simple Vector space Search Engine using Python

Amit Sethi

Vector Space model

- An Algebric model for representing text.
  - I am taking the dog for a walk .
  - Two dogs are walking in the park.

Resultant Vectors

```
[1,1]
[1,1,0]
```

Sample python code

- All my Brain post by Dennis
  - http://allmybrain.com/2007/10/19/similarity-of-texts-the-vector-space-model-with-python/

Removing STOP words

```
return [word for word in  list if word
        not in self.stopwords  ]
```

Stemming

- The Porter stemming algorithm (or ‘Porter stemmer’) is a process for removing the commoner morphological and inflexional endings from words in English.

Stemming Code

```
return [self.stemmer.stem(word,0,len(word)-1)
        for word in words]
```

Creating the Vectors

```
vector = [0] * len(self.vectorKeywordIndex)
wordList = self.parser.tokenise(wordString)
wordList = self.parser.removeStopWords(wordList)
for word in wordList:
        vector[self.vectorKeywordIndex[word]] += 1
        #Use simple Term Count Model
return vector
```

Adding numpy to the whole scenario

```
ratings = [numpy.cos(queryVector, documentVector)
           for documentVector in self.documentVectors]
```

Other things to explore

- NLTK
- Pylucene

Thank You!

Thank You Hope you find the tutorial useful