Natural Language Processing with Python

Group NLP

What is Natural Language Processing?

Natural language refers to the way language has evolved through everyday, casual use. It often breaks grammatical rules and even changes what some words mean.

An expert system needs to be trained to be able to make observations about any natural language text.

Our Project

Goal: Be able to predict the rating a movie has based on its movie description.

Process:

1. Data Processing
2. LDA
3. Sentiment Analysis

Data Processing

A series of operations on data, to retrieve, transform, or classify information into a format that can easily analyzed and processed by computers.

Data Processing

1

Tokenization

Text is split down into lowercase, unpunctuated words.

2

Words < 3 Removed

Words less than three words in length are ignored.

3

Stopwords Removed

Words that are the most commonly used in a language. (ex. “the”, “is” and “and”)

Data Preprocessing

4

Lemmatized

Words are converted to first-person variants and verbs are made present tense. (ex. “am”, “are”, “is” become “be”)

5

Stemmed

Words are reduced to their root form. (ex. “caresses” and “caress”)

Without Processing

	review	index
0	One of the other reviewers has mentioned that ...	0
1	A wonderful little production. The...	1
2	I thought this was a wonderful way to spend ti...	2
4	Petter Mattei's "Love in the Time of Money" is...	4
5	Probably my all-time favorite movie, a story o...	5

With Processing

```
0     [review, mention, watch, episod, hook, right, ...
1     [wonder, littl, product, film, techniqu, unass...
2     [think, wonder, spend, time, summer, weekend, ...
4     [petter, mattei, love, time, money, visual, st...
5     [probabl, time, favorit, movi, stori, selfless...
6     [sure, like, resurrect, date, seahunt, seri, t...
9     [like, origin, wrench, laughter, like, movi, y...
14    [fantast, movi, prison, famous, actor, georg, ...
16    [film, simpli, remak, film, fail, captur, flav...
18    [rememb, film, film, watch, cinema, pictur, da...
Name: review_processed, dtype: object
```

Latent Dirichlet Allocation (LDA)

A generative probabilistic model that represents data as sets belonging to various latent topics, where each topic is defined by the different probabilities of words.

LDA

A way of automatically discovering **topics** that the given documents contains.

For example:
“I like to eat broccoli and bananas.”
“Chinchillas and kittens are cute.”
“My sister adopted a kitten yesterday.”
“I ate a banana and spinach smoothie for breakfast.”
“Look at this cute hamster munching on broccoli.”

Topic A: 30% broccoli, 15% banana, 10% breakfast, 10% munching
Topic B: 20% chinchillas, 20% kittens, 20% cute, 15% hamster

Unsupervised Learning

This topic modeling algorithm can be referred to as a form of **unsupervised learning**.

Clustering is often used in machine learning to group together entities that are similar to uncover patterns.

LDA groups together similar words to uncover latent, or hidden, topics.

Sentiment Analysis

A type of data mining that measures the inclination of people’s opinions through natural language processing, computational linguistics and text analysis.

Sentiment Analysis Methods

Sentiment analysis uses various NLP methods and algorithms.

Machine learning techniques:

A sentiment analysis task is usually modeled as a classification problem.
A classifier is fed a text and returns a category, e.g. positive, negative, or neutral.

Bag of Words

A Bag of Words model refers to creating a dictionary from all unique words present in a text.

Extreme cases can be filtered out of the dictionary, such as those that appear in:

- less than 15 documents
- more than half of the documents
- after the above two steps, keep only the first 100,000 most frequent tokens

Libraries and Technologies Required

NLTK

Provides data sets and modules that support natural language processing.

NumPy, Scipy, Gensim

Gensim provides a fast implementation of LDA and requires NumPy and Scipy to work.

Jupyter

Provides data visualizations to make information easier to understand.

What will our demonstration be doing?

First, we will train a Bag of Words model using 50,000 movie reviews that are labeled either positive or negative.

We will use this model to classify movie descriptions into topics that our model has generated.

After classification, we will compare the weight of the negative and positive topics found in our movie description to predict the movie's rating.

Demonstration