DATA SCIENCE WITH PYTHON : TEXT ANALYTICS #1853

NAME : RICA GHOSH

SERIAL NO : 229

BATCH : 7

TEXT ANALYTICS

- Text analytics is the process of extracting information and uncovering actionable insights from unstructured text.
- Text analytics allows data scientists and analysts to evaluate content to determine its relevancy to a specific topic.
- Researchers mine and analyze text by leveraging sophisticated software developed by computer scientists.
- Text Analytics is also known as text mining or text data mining.

TEXT ANALYTICS PROCESS

1. Data Gathering
2. Data Preparation
3. Data Analysis
4. Data Visualization

1. DATA GATHERING

- Text analytics begins with collecting the text to be analyzed -- defining, selecting, acquiring, and storing raw data.
- This data can include text documents, web pages (blogs, news, etc.), and online reviews, among other sources.
- Data sources can be internal or external to an organization.

➢ **Internal Data** - This is the data you generate every day, from emails and chats, to surveys, customer queries, and customer support tickets. You just need to export it from your software or platform as a CSV or Excel file, or connect an API to retrieve it directly.

➢ **External Data** - This is text data about your brand or products from all over the web. You can use web scraping tools, APIs, and open datasets to collect external data from social media, news reports, online reviews, forums, and more, and analyze it with machine learning models.

2. DATA PREPARATION

- Once data is acquired, the enterprise must prepare it for analysis. The data must be in the proper form to work with machine learning models that will be used for data analysis.
- Stages in data preparation are :
- Text Cleansing
- Tokenization
- Part-of-Speech Tagging
- Parsing
- Stemming and lemmatization
- Stop word removal

TEXT CLEANSING :

- It removes any unnecessary or unwanted information, such as ads from web pages. Text data is restructured to ensure data can be read the same way across the system and to improve data integrity (also known as "text normalization").

TOKENIZATION :

- It breaks up a sequence of strings into pieces (such as words, keywords, phrases, symbols, and other elements) called tokens. Semantically meaningful pieces (such as words) will be used for analysis.

PART-OF-SPEECH TAGGING (POS) :

- It assigns a grammatical category to the identified tokens.

PARSING :

- It creates syntactic structures from the text based on the tokens and PoS models. Parsing algorithms consider the text's grammar for syntactic structuring. Sentences with the same meaning but different grammatical structures will result in different syntactic structures.

STEMMING AND LEMMATIZATION :

- It standardizes words by reducing them to their root forms.

STOP WORD REMOVAL :

- It filters out common words that add little or no unique information, for example, prepositions and articles (at, to, a, the).

3. DATA ANALYSIS

○ Data analysis is the process of analyzing the prepared text data.

○ Text Analysis Methods & Techniques :

- Text Classification
- Text Extraction
- Word Frequency
- Collocation
- Concordance
- Word Sense Disambiguation
- Clustering

TEXT CLASSIFICATION :

- Text classification is the process of assigning predefined tags or categories to unstructured text.
- It's considered one of the most useful natural language processing techniques because it's so versatile and can organize, structure, and categorize pretty much any form of text to deliver meaningful data and solve problems.
- The most common text classification tasks include :

➢ **Sentiment analysis** - It uses powerful machine learning algorithms to automatically read and classify for opinion polarity (positive, negative, neutral) and beyond, into the feelings and emotions of the writer, even context and sarcasm.

➢ **Topic modeling** - It automatically organizes text by subject or theme.

➢ **Intent detection** - It is often used to automatically understand the reason behind a particular feedback.

- There are two main algorithms that can be used to solve Text Classification problems:

- ➢**Rule-based Systems** - It rely on hand-crafted grammatical rules that need to be created by experts in linguistics, or knowledge engineers.

- ➢**Machine learning algorithms -** Machine learning models, on the other hand, are based on statistical methods and learn to perform tasks after being fed examples (training data). The most frequently used ml algorithms are :

- ▪ Naive Bayes (NB) family of algorithms
- ▪ Support Vector Machines (SVM)
- ▪ Deep learning algorithms

TEXT EXTRACTION :

- Text extraction extracts pieces of data that already exist within any given text.
- You can extract things like keywords, prices, company names, and product specifications from news reports, product reviews, and more.
- You can automatically populate spreadsheets with this data or perform extraction in concert with other text analysis techniques to categorize and extract data at the same time.
- The most common text extraction tasks include :

➢ **Keyword extraction** – It can be used to index data to be searched and to generate word clouds (a visual representation of text data).

➢ **Entity Recognition** - A named entity recognition (NER) extractor finds entities, which can be people, companies, or locations and exist within text data.

WORD FREQUENCY :

- Word frequency is a text analysis technique that measures the most frequently occurring words or concepts in a given text using the numerical statistic TF-IDF (term frequency-inverse document frequency).
- You might apply this technique to analyze the words or expressions customers use most frequently in support conversations.

COLLOCATION :

- Collocation helps identify words that commonly co-occur.
- Collocation can be helpful to identify hidden semantic structures and improve the granularity of the insights by counting bigrams (two adjacent words) and trigrams (three adjacent words) as one word.

CONCORDANCE :

- Concordance helps identify the context and instances of words or a set of words.
- It can also be used to decode the ambiguity of the human language to a certain extent, by looking at how words are used in different contexts, as well as being able to analyze more complex phrases.

WORD SENSE DISAMBIGUATION :

- It's very common for a word to have more than one meaning, which is why word sense disambiguation is a major challenge of natural language processing.
- Smart text analysis with word sense disambiguation can differentiate words that have more than one meaning, but only after training models to do so.

CLUSTERING :

- Text clusters are able to understand and group vast quantities of unstructured data.
- This is known as unsupervised machine learning.
- Google is a great example of how clustering works.
- Google's algorithm breaks down unstructured data from web pages and groups pages into clusters around a set of similar words or n-grams (all possible combinations of adjacent words or letters in a text).
- So, the pages from the cluster that contain a higher count of words or n-grams relevant to the search query will appear first within the results.

4. DATA VISUALIZATION

○ Visualization is the process of transforming analysis into actionable insights, representing the data in graphs, tables, and other easy-to-understand representations.

○ Data visualization boosts the value of the text mining results by transforming complex concepts into compelling and easily-to-understand visuals.

○ Some data visualization tools are:

- Tableau
- Looker
- Google Data Studio
- MonkeyLearn Studio

APPLICATIONS OF TEXT ANALYTICS

- Social Media Monitoring
- Brand Monitoring
- Customer Service
- Voice of Customer (VoC) & Customer Feedback
- Business Intelligence
- Sales and Marketing
- Product Analytics

ADVANTAGES OF TEXT ANALYSIS

- Helps identify the root of a problem (or source of satisfaction).
- Enables emerging trends to surface, that many feedback surveys limit or restrict.
- Issues can be prioritized quickly and efficiently.
- Customers' ideas and suggestions materialize, leading to an enhanced digital experience.

TEXT ANALYSIS TOOLS

OPEN SOURCE LIBRARIES

NLTK

SpaCy

Scikit-learn

TensorFlow

PyTorch

Keras

CoreNLP

OpenNLP

SAAS API

MonkeyLearn

Google Cloud NLP

IBM Watson

Lexalytics

MeaningCloud

Amazon Comprehend

Aylien

Clarabridge

RESOURCES

- https://monkeylearn.com/text-analysis/
- https://tdwi.org/articles/2019/06/03/adv-all-introduction-to-using-text-analytics-and-nlp.aspx
- https://www.lexalytics.com/lexablog/text-analytics-functions-explained
- https://mopinion.com/what-is-text-analytics-benefits/
- https://www.geeksforgeeks.org/text-mining-in-data-mining/
- https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

THANK YOU