09. Natural Language Analysis and Big Data Analysis

Telung Pan Ph.D.
telung@mac.com

Text Analysis

- Google Cloud Natural Language API reveals the structure and meaning of text by offering machine learning models.
- Extract information about people, places, events and much more, mentioned in text documents, news articles or blog posts.
- To understand sentiment about your product on social media or parse intent from customer conversations happening in a call center or a messaging app
- Analyze text uploaded storage on Google Cloud Storage.

Sentiment Analysis

Analyzing Sentiment

- Inspects the given text and identifies the prevailing emotional opinion within the text

- Especially to determine a writer's attitude as positive, negative, or neutral.

- lexicon-based techniques the neutrality score of the words is taken into account in order to either detect neutral opinions (Ding and Liu, 2008) or filter them out and enable algorithms to focus on words with positive and negative sentiment (Taboada et al, 2010).

- Statistical techniques are used, some researchers consider that the objective (neutral) sentences of the text are less informative and thus they filter them out and focus only on the subjective statements in order to improve the binary classification (Bo Pang and Lillian Lee, 2002).In other cases they use hierarchical classification where the neutrality is determined first and sentiment polarity is determined second (Wilson et al, 2005).

- In most academic papers of sentiment analysis that use statistical approaches, researchers tend to ignore the neutral category under the assumption that neutral texts lie near the boundary of the binary classifier.

- Koppel and Schler (2006) showed in their research both of the above assumptions are false. They suggested that as in every polarity problem three categories must be identified (positive, negative and neutral) and that the introduction of the neutral category can even improve the overall accuracy. Their work was primarily focused on SVM and they used geometric properties in order to improve the accuracy of their three binary classifiers.

- Chi-square provided better results for most classifiers.

- An intuitive explanation of why neutral class is important is the following: Not all things are black and white and not all sentences have a sentiment. How would you classify the sentence “the weather is hot”? Is it positive or negative?

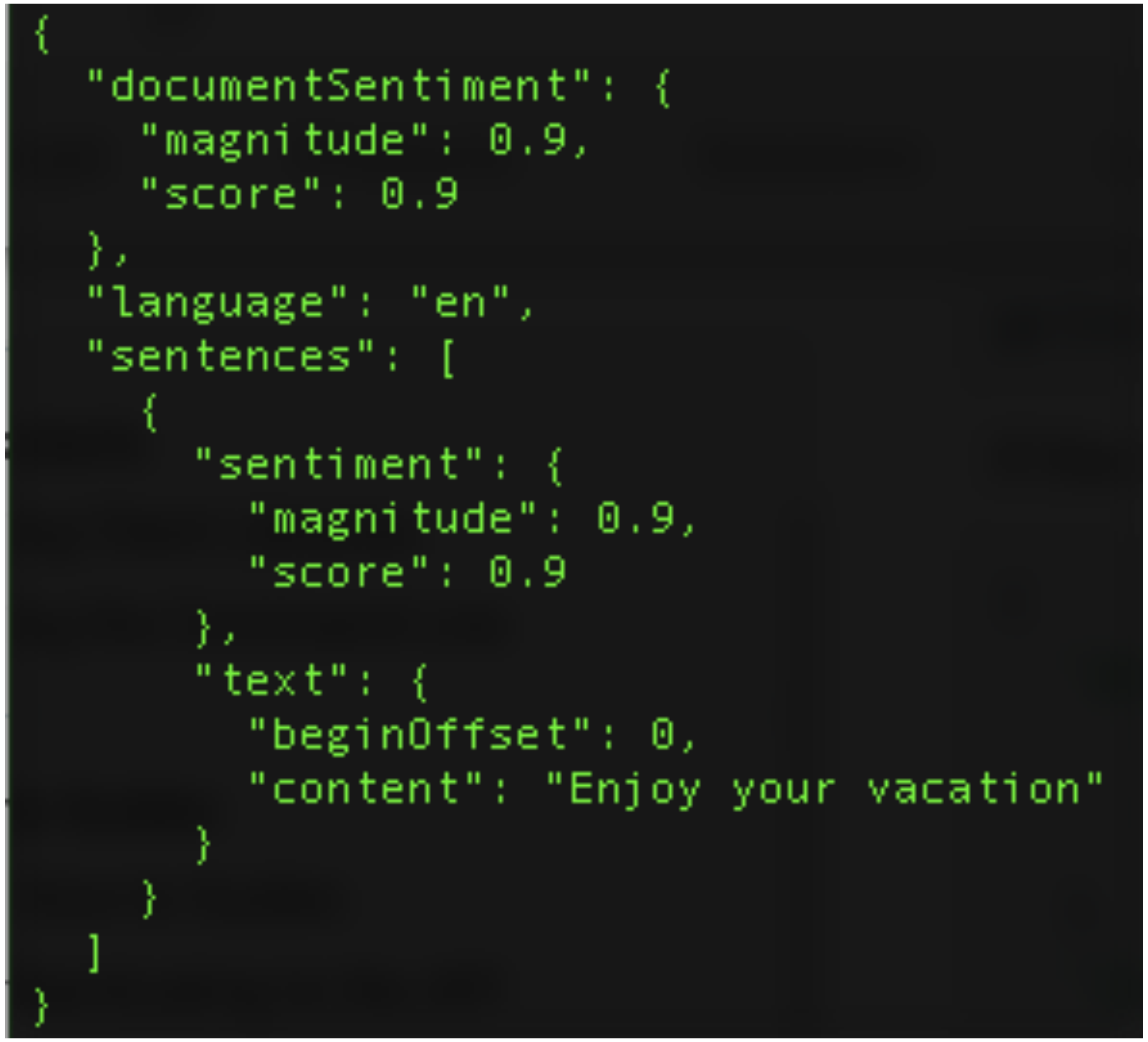
- When you use only 2 classes you basically force the features/words to be classified as either positive or negative leaving no room for neutrality. Professors Koppel and Schler published a paper about this called “The Importance of Neutral Examples for Learning Sentiment”

- Classifiers: 3 Naïve Bayes variations (Multinomial, Binarized and Bernoulli), Max Entropy, SVMs, Softmax Regression, Adaboost and more.

- Criteria: The overall accuracy, the variation across different datasets, the training and evaluation speed, their ability to parse large amount of data and the amount of resources that they use in terms of CPU and RAM.

- To perform sentiment analysis, use the gcloud command line tool and use the --content flag to identify the content to analyze:

gcloud ml language analyze-sentiment --content="Enjoy your vacation”



```
{
  "documentSentiment": {
    "magnitude": 0.9,
    "score": 0.9
  },
  "language": "en",
  "sentences": [
    {
      "sentiment": {
        "magnitude": 0.9,
        "score": 0.9
      },
      "text": {
        "beginOffset": 0,
        "content": "Enjoy your vacation"
      }
    }
  ]
}
```

- JSON format:
  - JavaScript Object Notation
  - 大括號表示物件
  - 中括號表示陣列
  - {“magnitude”: 0.9, “score”: 0.9}
- DocumentSentiment.score:
  Positive sentiment with a value greater than zero, and negative sentiment with a value less than zero.

Big Data - BigQuery

Google BigQuery

- BigQuery designed to focus on analyzing data to find meaningful insights using familiar SQL and you don't need a database administrator.
- BigQuery enables you to analyze all your data by creating a logical data warehouse over managed, columnar storage as well as data from object storage, and spreadsheets.
- BigQuery makes it easy to securely share insights within your organization and beyond as datasets, queries, spreadsheets and reports.
- BigQuery allows organizations to capture and analyze data in real-time using its powerful streaming ingestion capability so that your insights are always current.
- BigQuery is free for up to 1TB of data analyzed each month and 10GB of data stored.

bq Command

- gcloud auth login
- bq show
- bq show publicdata:samples.shakespeare

Run a query

-
  ```
  bq query "SELECT word, SUM(word_count) as count
  FROM publicdata:samples.shakespeare WHERE word
  CONTAINS 'raisin' GROUP BY word”
  ```

No Result

-

  ```
  bq query "SELECT word FROM publicdata:samples.shakespeare WHERE word = 'huzzah' IGNORE CASE"
  ```

巨量資料分析示範

- 資料來源：
  All names are from Social Security card applications for births that occurred in the United States after 1879.

- Open the file named yob2010.txt to see what it looks like. The file is a comma-separated value (CSV) file with the following three columns: name, sex (M or F), and number of children with that name. The file has no header row.

- Copy or move the yob2010.txt file into the directory you are using to run bq commands.

- Use the bq ls command to see whether your default project has any existing datasets.

  ```
  bq ls
  ```

- Run bq ls again to list the datasets in a specific project by including the project ID followed by a colon (:). The following example lists the datasets in the publicdata project.

  ```
  bq ls publicdata:
  ```

- Use the bq mk command to create a new dataset named babynames in your default project. A dataset name can be up to 1,024 characters long, and consist of A-Z, a-z, 0-9, and the underscore, but it cannot start with a number or underscore, or have spaces.

  ```
  bq mk babynames
  ```

  ```
  bq ls
  ```

- Upload the table
- The bq load command creates or updates a table and loads data in a single step.
- Run the bq load command to load your source file into a new table called names2010 in the babynames dataset you created above. By default, this runs synchronously, and will take a few seconds to complete.

  ```
  bq load babynames.names2010 yob2010.txt
  name:string,gender:string,count:integer
  ```

- The bq load command arguments:
  - datasetID: babynames
  - tableID: names2010
  - source: yob2010.txt
  - schema: name:string,gender:string,count:integer

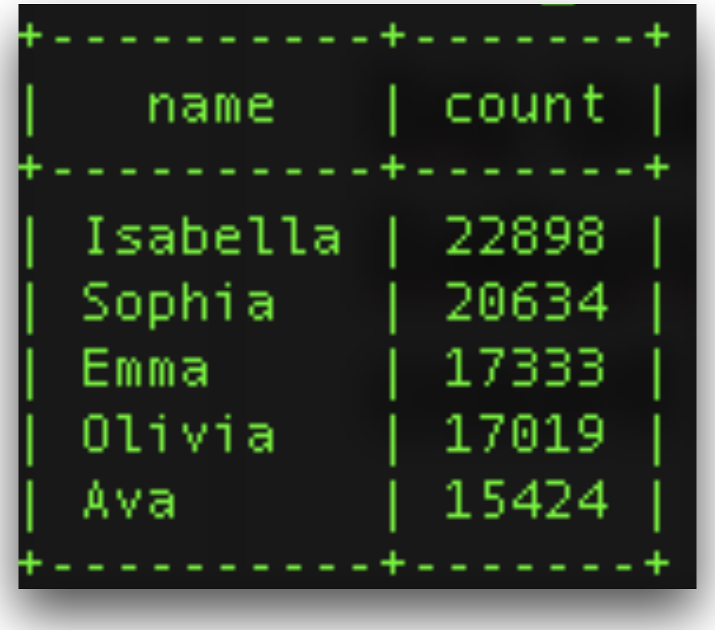
- Run bq ls to confirm that the table now appears in the dataset:

  ```
  bq ls babynames
  bq show babynames.names2010
  ```

Run queries

- Run the following command to return the most popular girls' names:

  bq query "SELECT name,count FROM babynames.names2010 WHERE gender = 'F' ORDER BY count DESC LIMIT 5”



```
+-----------+-------+
|   name    | count |
+-----------+-------+
| Isabella  | 22898 |
| Sophia    | 20634 |
| Emma      | 17333 |
| Olivia    | 17019 |
| Ava       | 15424 |
+-----------+-------+
```

- Run the following command to see the most unusual boys' names. The minimum count is 5 because the source data omits names with fewer than 5 occurrences.

  ```
  bq query "SELECT name,count FROM
  babynames.names2010 WHERE gender = 'M' ORDER BY
  count ASC LIMIT 5"
  ```

Clean up

- To avoid incurring charges to your Google Cloud Platform account for the resources used in this quickstart:

- Run the bq rm command to remove the babynames dataset. Use the -r flag to delete all tables in the dataset, include the names2010 table.

  ```
  bq rm -r babynames
  ```

- Confirm the delete command by typing y.