Project D7: KAGGLE - NATURAL LANGUAGE PROCESSING WITH DISASTER TWEETS

Project repository: <https://github.com/BelindaLep/IDS_project>

TEAM:
Marilin Ahvenainen
Belinda Lepmets

Business understanding

- **Identifying your business goals**
  - **Background** - As Twitter has become an important communication channel in times of emergency, then different agencies (i.e. disaster relief organizations and news agencies) are interested in programmatically monitoring Twitter. Unfortunately it's not always clear whether a person's words are actually announcing a disaster.
  - **Business goals** - To be able to monitor emergencies via Twitter more efficiently using a machine learning model that is able to predict which Tweets are about real disasters and which one's aren't.
  - **Business success criteria** - Success will be evaluated using F1 between the predicted and expected answers. Sample submission should include id and target (prediction).
- **Assessing your situation**
  - **Inventory of resources** -
    data: test.csv - test dataset
    train.csv - training dataset
    sample_submission.csv
    software: colab notebook
  - **Requirements, assumptions, and constraints** - building a machine learning model by 7th of December, poster submission deadline: Monday, 12th of December, at noon (12:00)
  - **Risks and contingencies** - internet outage - could use wifi at school or library. Colab notebook is not working - using alternative notebook.
  - **Terminology** :
    word embedding - a term used for the representation of words for text analysis, typically in the form of a real-valued vector that encodes the meaning of the word such that the words that are closer in the vector space are expected to be similar in meaning.

    F1 score- can be interpreted as a harmonic mean of precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

    $$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

    Precision- Precision is defined as the ratio of correctly classified positive samples (True Positive) to a total number of classified positive samples (either correctly or incorrectly).

$$precision = \frac{TP}{TP + FP}$$

Recall- Recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected. $$recall = \frac{TP}{TP + FN}$$

True Positive [TP] = your prediction is 1, and the ground truth is also 1
False Positive [FP] = your prediction is 1, and the ground truth is 0
False Negative [FN] = your prediction is 0, and the ground truth is 1

- **Costs and benefits** - Not meant to benefit a specific business.

- **Defining your data-mining goals**
  - **Data-mining goals** - choosing word embedding technique, data visualization for poster
  - **Data-mining success criteria** - minimum model accuracy: 80%

Data understanding

- **Gathering data**
  - **Outline data requirements**
    - tabular data type consisting of : id, keyword (nominal), location (nominal), text (nominal), target (nominal)
    - Id should be an integer, keyword should be a string, location should also be a string, text should be a string too and target should be an integer
  - **Verify data availability** - required data exists and is available from the competitions page after joining the competition.
  - **Define selection criteria**
    - Id columns values range from 1 to 10873 in the train dataset and from 1 to 10875 in the test dataset, as this is how many rows of data they have.
    - Keyword does not have a specific range - a lot of disaster-related words.
    - Location does not have a specific range - you can tweet from basically anywhere.
    - Text does not have a specific range, but can not be over 280 characters as this is the limit for tweets.
    - Target has two values 0 or 1 (only in train dataset)
- **Describing data -** we will use 2 csv format datafiles:
  - one for training the prediction model (train.csv), it consists of 10873 tweets. Dataset consist of 5 columns, an ‘id’ column that is for a unique identifier of each tweet, a ‘text’ column that is for a text of the tweet, a ‘location’ column that is for a location the tweet was sent from (may be blank), a ‘keyword’ column that is for a particular keyword from the tweet like ‘tornado’, ‘earthquake’, ‘blizzard’ or something like that (may be blank), a ‘target’ column that denotes whether a tweet is about a real disaster (1) or not (0)
  - Second for testing the prediction model (test.csv) it consists of 10875 tweets, it consist of 4 columns, an ‘id’ column that is for a unique identifier of each tweet, a ‘text’ column that is for a text of the tweet, a ‘location’ column that is for a location the tweet

was sent from (may be blank), a ‘keyword’ column that is for a particular keyword from the tweet like ‘tornado’, ‘earthquake’, ‘blizzard’ or something like that (may be blank)

- **Exploring data -** ‘id’ and ‘text’ columns are consistent and no rows are empty (also ‘target’ column in train dataset). ‘Location’ and ‘keyword’ columns are inconsistent; they can have empty rows, location can be a very wide area or not a real place at all.
- **Verifying data quality -** Looks like this data is in the right format for us, consists of all the necessary columns that we think we will need, it is accessible and should be good enough to achieve our goals.

Planning your project

Tasks,methods,tools:

1. Getting familiar with the data provided (test and train datasets) (0.5h per member)
2. Getting familiar with word embedding techniques (for example Word2Vec, GloVe, BERT, TF-IDF) and choosing a preferable word embedding technique. (4h per member)
3. Building a machine learning model that predicts which Tweets are about real disasters and which one’s aren’t using Colab Notebook. (20h per member)
4. Calculating precision and recall and measuring the F1 score for the model (1h per member)
5. Improving the model if necessary, if the model accuracy is less than 80% (2h per member)
6. Creating a poster for the poster session (including visualizing the data) (4h per member)