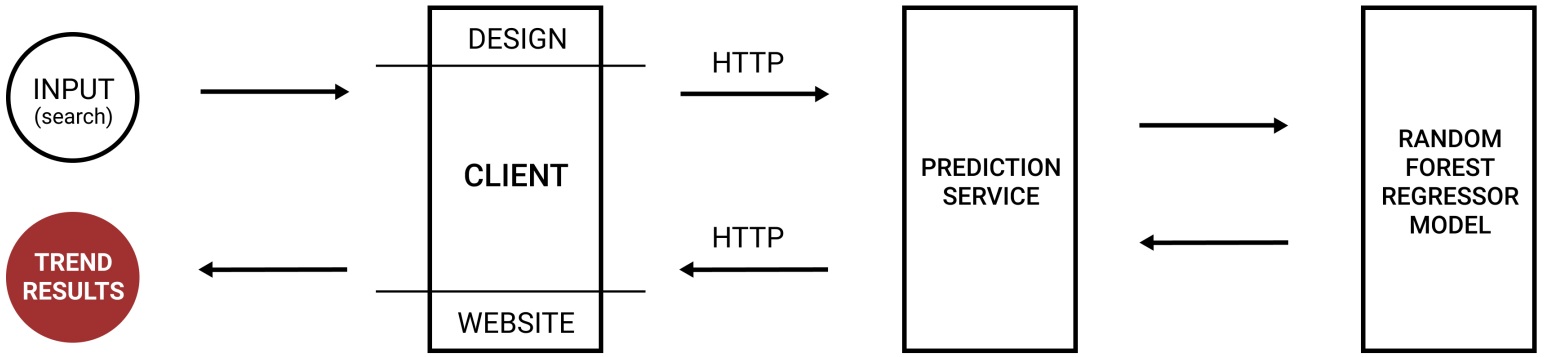
YouTube Viral Predictor

Team:
Rainey Chak
Suofei Feng (Presenter)
Yen-Ming Huang
John Solo

Product Overview



Data Science Component

- Original Input:
  - Features: title, channel title, category id, publish time, tags, description, region
  - 147k in train set, 7.7k in test set
- Processing:
  - Add ‘region’ to predictor features
  - Word embedding using FastText model:
    - title, channel title, tags, description
  - Normalize number of views as the target variable
- Output:
  - True value: normalized values of ‘views’
  - Predicted value: trend score

More about the model

- Random Forest Regressor
- Grid search tuning on:
  - number of estimators
  - max depth of each tree
- Evaluation Metric: root mean squared error

Next steps

PREDICT HOW LONG TO TREND

- Use the date difference between the trending date and the publish date as the label

SORT THROUGH DIFFERENT CATEGORIES

- Given a key word, return a list of sub-categories ranked according to the predicted popularity

EXPAND TO SUPPORT DIFFERENT LANGUAGES

- Train on different language corpus using embedding models for different languages

Thanks!

Oct. 25, 2019

SAP, Palo Alto