Language Detection & Processing

NLP580 Hackathon Group 5

April Chung, Kate Bosshart, Kendra Gedney, RD Mitra, & Sooeun Oh

Language Identification

```
Pernyataan Umum tentang Hak-Hak Asasi Manusia      Mukadimah      Menimbang
bahwa pengakuan atas martabat alamiah dan hak-hak yang sama dan mutlak dari
semua anggota keluarga manusia adalah dasar kemerdekaan, keadilan dan
```

Excerpt of Language from Corpus 2

Apply `udhr_identify.py` code, based on HW3 with a subset of languages

```
english     0.546937
spanish     0.650000
Kaonde      0.559289
Arab        0.279412
romani      0.679221
oromo       0.736307
Hani        0.324111
Indonesian  0.997036
The best language guess for the unknown is Indonesian
```

Based on a 0.997 Spearman's rank correlation, we are confident the language is Indonesian!

After applying language detectors (Polyglot & Spacy) to various documents in our corpus, the majority of methods detected language as **Indonesian**

- Some documents in our corpus were detected as **Malay**. These languages are closely related, with similar words

Named Entity Recognition (NER)

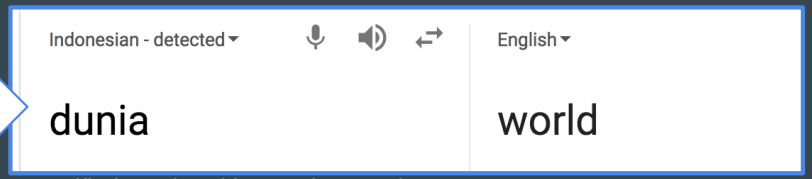
- For future research, we are planning to apply Named Entity Recognition (NER) to identify people and places, to help further substantiate that the language is Indonesian.
  - We make the assumption that the Entities named in our corpus would align with the countries in which Indonesian is spoken.

Part of Speech (PoS) Tagging

- **Use:** NLTK package and ‘stopwords_l2.txt’

```
('dunia', 'NN'),
```

NLTK indicates ‘dunia’ is a noun



Verified results with Google Translate

```
('hidup', 'NN'), ('kemerdekaan', 'NN'),
('luas', 'NN'), ('menimbang', 'NN'),
('negaranegara', 'NN'), ('anggota', 'NN'),
('berjanji', 'NN'), ('mencapai', 'NN'),
('kemajuan', 'JJ'), ('penghargaan', 'NN'),
('penghormatan', 'NN'), ('hakhak', 'NN'),
('asasi', 'NN'), ('manusia', 'NN'),
('kebebasankebebasan', 'NNP'), ('asasi',
'VBZ'), ('bekerjasama', 'NN'),
('perserikatan', 'NN'), ('bangsabangsa',
'NN'), ('menimbang', 'NN'), ('pengertian',
'JJ'), ('hakhak', 'NN'),
('kebebasankebebasan', 'VBD'),
```

Excerpt of NLTK PoS Tagging

- By comparing our PoS Tagging with the known structure of Indonesian sentences, we can further validate that the unknown language sample is Indonesian.
- Based on our research, Indonesian sentence word order is similar to English, using a **subject + verb + object** pattern.

Elasticsearch Index

We created an elasticsearch index that represents our test data documents and includes the result of your analytics

- Moved `elasticTest.py` to our dev folder, edited using vi to create a new index for our folder
- curl -X GET 'http://localhost:9200/mydocs/text_samples/1'

```
[kg729@vps 5_anly580]$ vi elasticTest.py
[kg729@vps 5_anly580]$ python elasticTest.py

My first review:
{'title': 'UDHR', 'language_id': 'id', 'language': 'indonesian', 'timestamp': '2018-11-26T20:51:55.280342'}

My search result:
{'took': 1, 'timed_out': False, '_shards': {'total': 5, 'successful': 5, 'skipped': 0, 'failed': 0}, 'hits': {'total': 1, '
max_score': 0.2876821, 'hits': [{'_index': 'mydocs', '_type': 'text_samples', '_id': '1', '_score': 0.2876821, '_source': {
'title': 'UDHR', 'language_id': 'id', 'language': 'indonesian', 'timestamp': '2018-11-26T20:51:55.280342'}}]}}
[kg729@vps 5_anly580]$
```