Language Detection using LDA

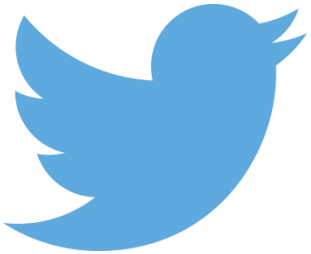
Can Taylan SARI
Volkan KÜÇÜK

Introduction

- The aim of this project is to create a model to infer language of a tweet by using LDA.
- Tweets are short and noisy,
- Language identifier resides inside of a tweet, which provides ground-truth data.

Main Objects of Twitter

- Users, → tweet, follow etc. (anyone or anything)
- Entities, → #hashtags, media, url, @user mentions
- Places, → geographic place, attached to a tweet
- Entities in Objects → As discussed.
- Tweets.



Structure of a Tweet

- Called as “status update”
- 140 character long
- Everthing of Twitter
- Main fields of a tweet;
  - id
  - coordinates
  - created_at
  - entities,
  - lang
  - Place
  - retweeted,
  - Text
  - User

Structure of a Tweet

Dataset Creation

- Twitter has a streaming API,
- With a user created twitter application, some key pairs are obtained,
- By using these keys a programmer/application can easily consume tweets.
- Twitter4j library to consume streaming API,
- Twitter4J is an unofficial Java library for the Twitter API.

Data Preprocessing

- As stated before tweets are noisy,
- To create a model first we have to remove all entities and unrelated data from each tweet,

Before : #facebook bored_rose @RayPyngotes what's up fat Clarkey unable to find a FB account https://t.co/u2I11wQvH3

After : what s up fat Clarkey unable to find a FB account

Data Preprocessing

- Lower case *:
  - All tweets are casted to lowercase.
- Remove punctuation :
  - All punctuation marks are removed.
- Remove whitespaces :
  - There must be only one space between words.
- Remove stop words *
- Stamming *

* will be implemented.

Collected Dataset

- Collected Data;
  - 2000 Spanish tweets,
  - 1996 English tweets,
  - 2000 Russian tweets,
  - 1980 Turkish tweets,
  - 2000 French tweets,
- Timestamp information is also collected.
- Tweets were collected March 26/27
- More tweets can be collected in order to create better model

Model

- We assume that;
  - Every tweet is a document,
  - Collected tweets are our corpus,

Sample Run

- Parameters
  - Number of tweets: ~10000
  - Number of languages: 5
  - Number of topics: 10
  - Number of most probable words shown in topic distributions: 10

Dataset

Garantili TT dk da TT yapıyoruz

Göksu sabahtan beri hangisini alayım diyor Çıldırcam hepsi aynı değil mi bunların

ardahan press sayfası olarak herkese hayırlı cumalar diliyoruz aylık değil haftalık degil günlük değil saatlik

Спартак Чемпион

Ненавижу Жизнь

мне нравитсяТакой стих готов даже выучить и написать в тетрадочку любимых стихов

Escargophone trouv Promesse une baleine Nous nous reverrons ici

On rigole trop avec moi pck jui trop drle la vrit

Build your team for Fantasy Football on Mondo Goal

Example Topics

Topic 4th:	
est	0.022
le	0.021
pas	0.018
je	0.016
c	0.014
et	0.014
j	0.014
l	0.014
les	0.013
une	0.012

Topic 3th:	
I	0.032
the	0.027
To	0.026
you	0.019
and	0.018
in	0.017
t	0.014
of	0.013
for	0.013
s	0.013

Topic 8th:	
в	0.037
и	0.024
не	0.020
на	0.017
я	0.012
с	0.012
что	0.010
В	0.0078
а	0.0062
это	0.0059

Example Topics

Topic 2th:	
bir	0.0141
ve	0.0123
bu	0.0103
The	0.0103
da	0.0102
o	0.0080
from	0.0078
ne	0.0073
çok	0.0065
ya	0.0062

Topic 0th:	
ve	0.2405
via	0.0064
video	0.0029
gsdg	0.0025
unfollowers	0.0012
dan	0.0011
Check	0.0011
by	0.0010
seguidores	0.0010
yani	0.0010

Future Work

- Preprocess
  - To lower case, remove stopword, stemming etc.
- Optimize parameters
  - Number of topics, hyperparameters of model
- Evaluation of model
  - Quantitative (perplexity)
  - Qualitative