

Language Detection using LDA

Can Taylan SARI

Volkan KÜÇÜK

Outline

- Introduction
- Experiment I
 - Dataset
 - Model
 - Results (Topic Distributions)
- Experiment II & III
 - Dataset
 - Model
 - Workflow
 - Results (Topic Distributions & Inference)

Introduction

- The aim of this project is to create a model which detects the language of a tweet by using LDA.
- We assume that;
 - Every tweet is a document,
 - Collected tweets are our corpus,

Experiment I

Dataset

- 2000 tweets in 5 languages = 10000 tweets.
 - Turkish
 - Russian
 - English
 - French
 - Spanish
- Not preprocessed.

Model

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

Results (Topic distributions)

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:

l	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

Results (Topic distributions)

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

Experiment II

Dataset

- Tweets were collected in April 22/24
- Total ~75000 tweets (documents)
 - 13657 Spanish tweets
 - 20012 English tweets
 - 13266 Russian tweets
 - 13557 Turkish tweets
 - 14848 French tweets
- Total ~650000 words

Preprocessing

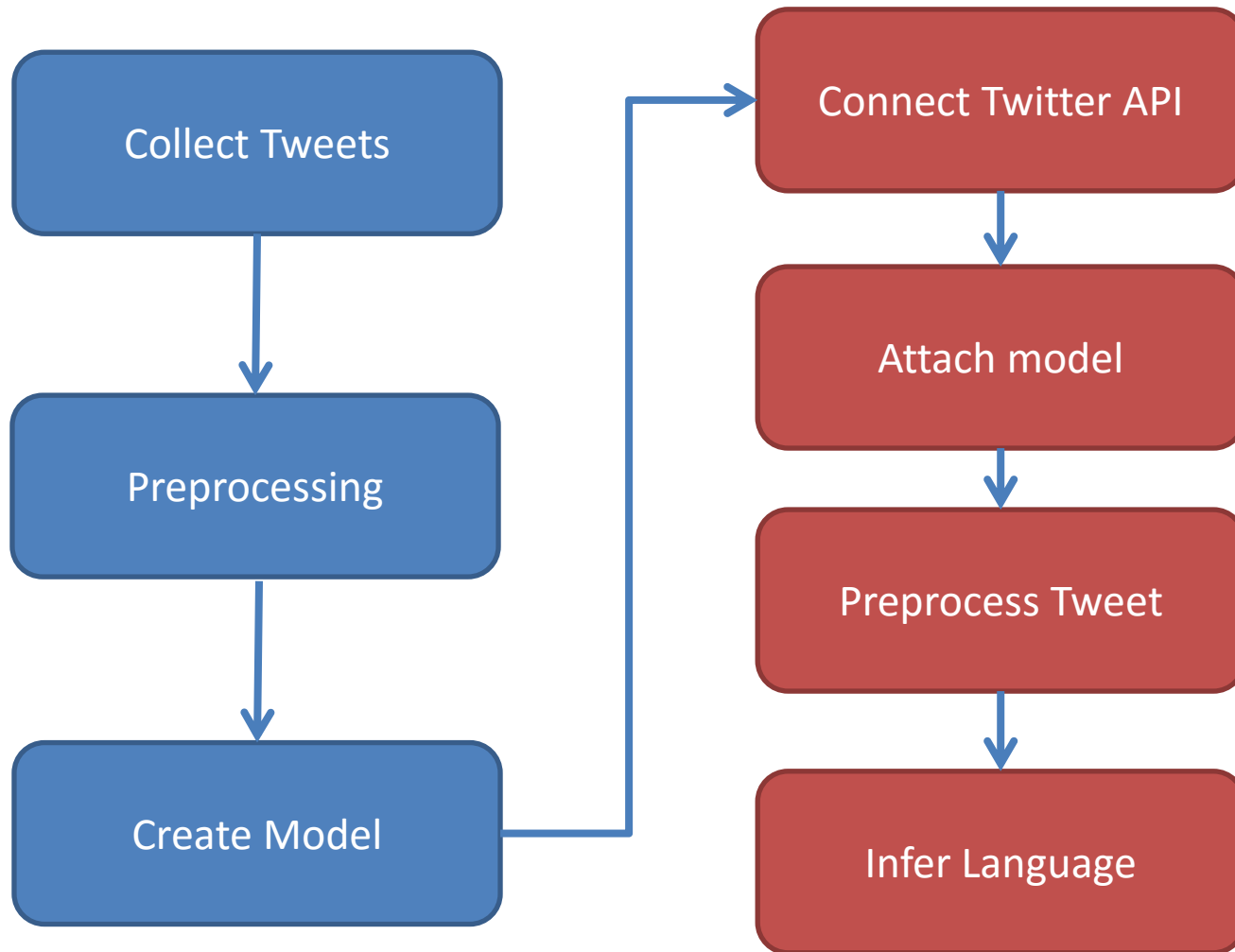
- ~~Remove punctuation :~~
 - All punctuation marks are removed.
- ~~Remove whitespaces :~~
 - There must be only one space between words.
- Lower case :
 - All tweets are casted to lowercase.
- Remove stop words
- Stemming *

* will be implemented.

Model

- Parameters
 - Number of tweets: ~75000
 - Number of languages: 5
 - Number of topics: 5
 - Number of most probable words shown in topic distributions: 20

Workflow



Results (Topic Distributions)

Topic 0th:

weather	0.0098
get	0.0085
channel	0.0049
now	0.0048
updates	0.0048
new	0.0048
just	0.0046
like	0.0039
via	0.0036
love	0.0036
one	0.0035
april	0.0035
can	0.0034
day	0.0032
pm	0.0028
good	0.0028
re	0.0028
will	0.0027
go	0.0026
video	0.0026

Topic 1th:

une	0.0100
ai	0.0074
plus	0.0057
qu	0.0054
moi	0.0037
suis	0.0037
faire	0.0028
via	0.0027
aime	0.0024
toi	0.0022
ah	0.0022
mme	0.0021
photo	0.0021
facebook	0.0021
mdr	0.0020
tre	0.0020
non	0.0019
сможет	0.0019
золотых	0.0019
собрал	0.0019

Topic 2th:

se	0.0126
te	0.0099
del	0.0083
mi	0.0077
al	0.0050
ms	0.0044
hoy	0.0035
este	0.0023
hay	0.0022
gracias	0.0021
feliz	0.0019
todos	0.0018
rt	0.0018
mejor	0.0018
quiero	0.0018
eso	0.0017
ver	0.0016
das	0.0016
mas	0.0016
vida	0.0015

Results (Topic Distributions)

Topic 4th:

не	0.0183
на	0.0168
мне	0.0046
онити	0.0043
но	0.0029
он	0.0020
новости	0.0018
голосдети	0.0018
видео	0.0014
хочу	0.0014
россии	0.0014
android	0.0012
они	0.0012
checked	0.0010
automatically	0.0009
followed	0.0009
она	0.0009
ios	0.0008
вообще	0.0008
unfollowed	0.0008

Topic 5th:

var	0.0034
kadar	0.0034
bi	0.0033
gibi	0.0030
olsun	0.0027
yok	0.0026
güzel	0.0024
iyi	0.0024
beşiktaş	0.0018
bile	0.0018
allah	0.0018
değil	0.0018
nisan	0.0017
türkiye	0.0016
sana	0.0016
sonra	0.0016
istanbul	0.0015
gol	0.0015
unfollowers	0.0014
abi	0.0014

Results (Inference)

- Attached our model to Twitter Stream API
- Collect **10000** (2000×5 languages) tweets
- Classify incoming tweets near **real time**

Results (Inference)

Language	# of Tweets	Misclassified	Accuracy
English	2000	179	% 91
French	2000	412	%79,4
Russian	2000	583	%70,85
Spanish	2000	423	%79,15
Turkish	2000	559	%72,05
Overall	10000	2156	%78,44

Results (Inference)

Original Tweet	Original	Inferred
@FoxNewsSunday @FoxNews Chris Wallace you sound like a jackass.Stop with the stuttering & trying to trip people up you sound ridiculous	English	English
My girl is the most beautiful girl in this planet ??	English	English
Coca Cola Coke Address Telephone note pad Tin 1987 https://t.co/B1nGyORKSN	English	French
@AnxietyXena non. Je suis pas ultra fan.	French	English
Şile bezi yazın en sevileni???? #moz #stil #style #styling #stylish #outfit #todaysoutfit #todaysook #lookoftheday	Turkish	English
İnsanlık, yeni bir doğuma gebe... https://t.co/Ghpw7VGGNO	Turkish	Turkish
#YoConAndalucía Rosa Aguilar preside en Córdoba la celebración del Día Internacional del Libro: La... https://t.co/d3NXEWDECI @susanadiaz	Spanish	Spanish
#Красноярск На Пасху в Москве будут "разгонять облака"... https://t.co/8oLlzQdILf	Russian	Russian
Правительство обсудит индексацию пенсий работающим пенсионерам в конце года...	Russian	Spanish

Experiment III

Dataset

- Tweets were collected in April 22/24
- Total ~75000 tweets (documents)
 - 13657 Spanish tweets
 - 20012 English tweets
 - 13266 Russian tweets
 - 13557 Turkish tweets
 - 14848 French tweets
- Total ~650000 words

Model

- Parameters
 - Number of tweets: ~75000
 - Number of languages: 5
 - Number of topics: 20
 - Number of most probable words shown in topic distributions: 20

Results (Topic distributions)

Topic 9th:

olsun	0.005
nisan	0.003
çocuk	0.002
istanbul	0.0025
türkiye	0.0020
kutlu	0.0020
vos	0.0012
izmir	0.0011
fire	0.0011
ulusal	0.009
İnsanlar	0.009
egemenlik	0.008
bayramı	0.008
bak	0.007
girls	0.007
pk	0.005
clip	0.004
friday	0.004
aracılığıyla	0.004
Atatürk	0.004

Topic 10th:

via	0.0146
prince	0.0056
direct	0.0014
music	0.0012
français	0.0010
rain	0.0010
purple	0.009
news	0.009
xd	0.008
france	0.008
french	0.008
mm	0.006
Mr	0.006
daily	0.006
wow	0.005
wind	0.005
found	0.005
stories	0.005
excuse	0.005
помощью	0.005

Topic 16th:

real	0.0029
madrid	0.0022
away	0.0016
vs	0.009
liga	0.006
god	0.006
nice	0.006
buried	0.005
health	0.005
style	0.005
san	0.005
love	0.005
ligne	0.004
good	0.004
also	0.004
high	0.003
just	0.003
bad	0.003
speed	0.003
ps	0.003

Questions/Answers