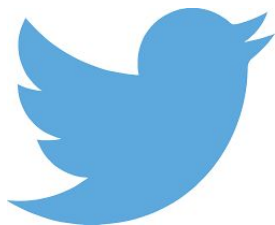


Analysing Prestigiousness of Universities

Jaeseok Huh, Moonyeong Lee

School of Computing

Our Proposal



1	Data Retrieval & Pre-processing	<ul style="list-style-type: none">• Twitter - One of public datasets• Rankings - via crawling
2	LDA with word-vectors	<ul style="list-style-type: none">• LIWC or the like
3	Ordered logistic regression	<ul style="list-style-type: none">• "Ordered" - to reward close estimation
4	Deviance Analysis	<ul style="list-style-type: none">• Explanatory analysis



@gabekwong is **visiting** our campus (**KAIST**) tomorrow for a **talk** on engineering proteases! **Excited** for this!

What are the *linguistic features* involved with *prestigious universities*?

Stage I.



Data Retrieval & Preprocessing

I. Dataset Summary

- **Ranking**

-  **55** out of top 60 universities
- excl. undistinguishable names like “*Washington University*”

- **Corpus**

-  **17.4K** out of 204K articles from 18 American Publications in 2013-2018
-  **5.1K** out of 1.8M posts from online boards

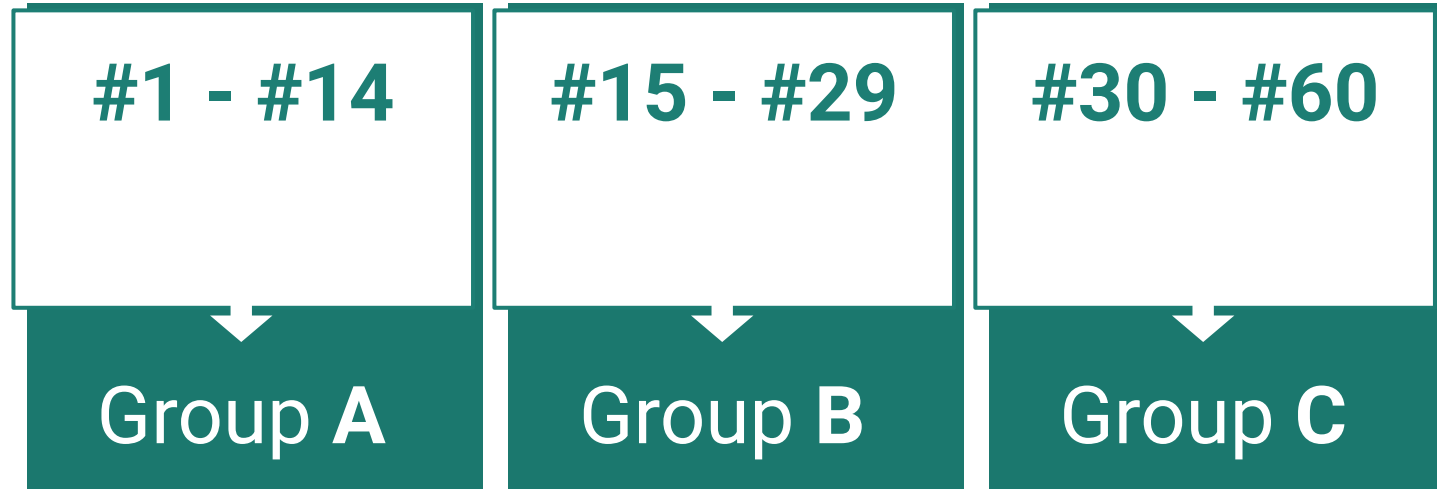
Crafting Regular Expression

University Name	Abbreviation	Where Condition
Princeton University		PRAGMA univ(Princeton)
Harvard University		PRAGMA univ(Harvard)
Yale University		PRAGMA univ(Yale)
University of Chicago		PRAGMA univ(Chicago)
Columbia University		PRAGMA univ(Columbia)
Massachusetts Institute of Technology	MIT	PRAGMA include(Massachusetts Institute of Technology) or PRAGMA abbr(MIT)
Stanford University		PRAGMA univ(Stanford)
University of Pennsylvania		PRAGMA univ(Pennsylvania) or PRAGMA univ(Denn)
Duke University		PRAGMA univ(Duke)
Johns Hopkins University		PRAGMA univ(Johns Hopkins) or PRAGMA univ(JohnsHopkins)
Northwestern University		PRAGMA univ(Northwestern)
California Institute of Technology	<u>Caltech</u>	PRAGMA include(California Institute of Technology) or PRAGMA include(caltech)
Dartmouth College		PRAGMA univ(Dartmouth)
Brown University		PRAGMA univ(Brown)
Vanderbilt University		PRAGMA univ(Vanderbilt)
Cornell University		PRAGMA univ(Cornell)
Rice University		PRAGMA univ(Rice)
University of <u>Notre</u> Dame		PRAGMA univ(Notre Dame)
Washington University in St. Louis		PRAGMA univ(Washington) and PRAGMA notuniv(Central Washington) and PRAGMA notuniv(Eastern Washington) and PRAGMA not univ(Georgia Washington)

Example of Regular Expressions

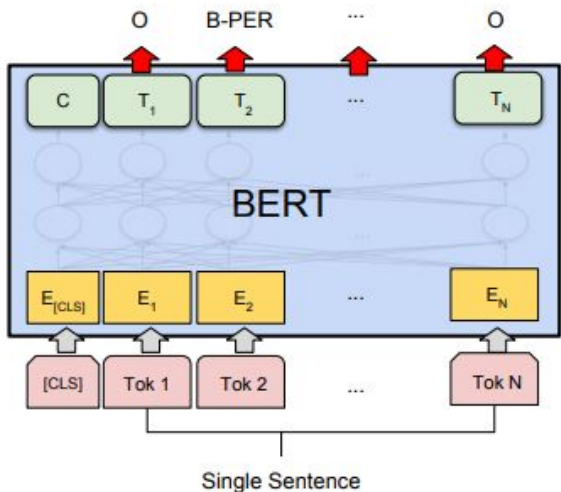
University Name	Regex	2019 Ranking
Princeton University	<code>((?:University of Princeton))(?:Princeton U...</code>	1
Harvard University	<code>((?:University of Harvard))(?:Harvard Unive...</code>	2
Yale University	<code>((?:University of Yale))(?:Yale University)))</code>	3
University of Chicago	<code>((?:University of Chicago))(?:Chicago Unive...</code>	3
Columbia University	<code>((?:University of Columbia))(?:Columbia Uni...</code>	3
Massachusetts Institute of Technology	<code>((?:Massachusetts Institute of Technology))(\...</code>	3
Stanford University	<code>((?:University of Stanford))(?:Stanford Uni...</code>	7
University of Pennsylvania	<code>((?:University of Pennsylvania))(?:Pennsylv...</code>	8
Duke University	<code>((?:University of Duke))(?:Duke University)))</code>	8
Johns Hopkins University	<code>((?:University of Johns.Hopkins))(?:Johns.H...</code>	10
Northwestern University	<code>((?:University of Northwestern))(?:Northwes...</code>	10

Grouping into Three



To focus on **general attributes**
rather than those of *specific* universities

Entity Detection: BERT-NER-classifier



(d) Single Sentence Tagging Tasks:
CoNLL-2003 NER

accuracy:	98.15%	precision:	90.61%	recall:	88.85%	FB1:	89.72
LOC:	precision:	91.93%	recall:	91.79%	FB1:	91.86	1387
MISC:	precision:	83.83%	recall:	78.43%	FB1:	81.04	668
ORG:	precision:	87.83%	recall:	85.18%	FB1:	86.48	1191
PER:	precision:	95.19%	recall:	94.83%	FB1:	95.01	1311

CoNLL-2003 Dataset

Stage II.
f(sentences) -> vector

GloVe, EMPATH,
and n-grams by tf-idf

GloVe

- Aggregated co-occurrence statistics from a corpus

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}

- The resulting representations showcase interesting **linear substructures** of the word vector space.

GloVe: Vector Arithmetic

find_closest_word (usa_school - "USA" + nationality)

USA School	mit	stanford	caltech	dartmouth	vanderbilt
JP School	tsinghua	japan	komura	waseda	hashimoto

find_closest_word (asian_school - nationality + "USA")

Asian Schools	kaist	snu	yonsei	sungkyunkwan	hanyang	postech	waseda	nus	ntu
Nationality	korea	korea	korea	korea	korea	korea	japan	singapore	singapore
USA Schools	usp	sdu	mannes	poynter	devry	dtic	biola	nccaa	uw


find_closest_word ("MIT" - "technology" + literature)

Literature	art	law	history
School	yale	yale	yale

Word Embeddings - EMPATH (instead of LIWC)

```
lexicon.analyze("he hit the other person", normalize=True)

# => {'help': 0.0, 'office': 0.0, 'violence': 0.2, 'dance': 0.0, 'money': 0.0, 'wedding': 0.0,
'valuable': 0.0, 'domestic_work': 0.0, 'sleep': 0.0, 'medical_emergency': 0.0, 'cold': 0.0, 'hate':
0.0, 'cheerfulness': 0.0, 'aggression': 0.0, 'occupation': 0.0, 'envy': 0.0, 'anticipation': 0.0,
'family': 0.0, 'crime': 0.0, 'attractive': 0.0, 'masculine': 0.0, 'prison': 0.0, ... }
```



```
Affective - "anger", "sadness", "swearing_terms" ...
Education - "school", "college", "business", "programming"...
Supporting - "help", "money", "vacation", "health"...
Interpersonal - "trust", "listen", "friends", "family"...
Working - "work", "office", "white_collar_job"...
Reputation - "occupation", "pride", "dispute", "royalty"...
Anti-social - "violence", "fight", "injury", "rage" ..
```

tf-idf (term frequency–inverse document frequency)

- A numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. (*Rajaraman, 2011*)
- In a nutshell, “**awesome**” or “**brown**” is much more important than “**the**”.

Stage III.

Machine Learning

Stage IV.

Analysis

Research Questions

1. Is it possible to classify which group a university falls into?
2. What are the linguistic features that distinguish high- and low-ranked universities?

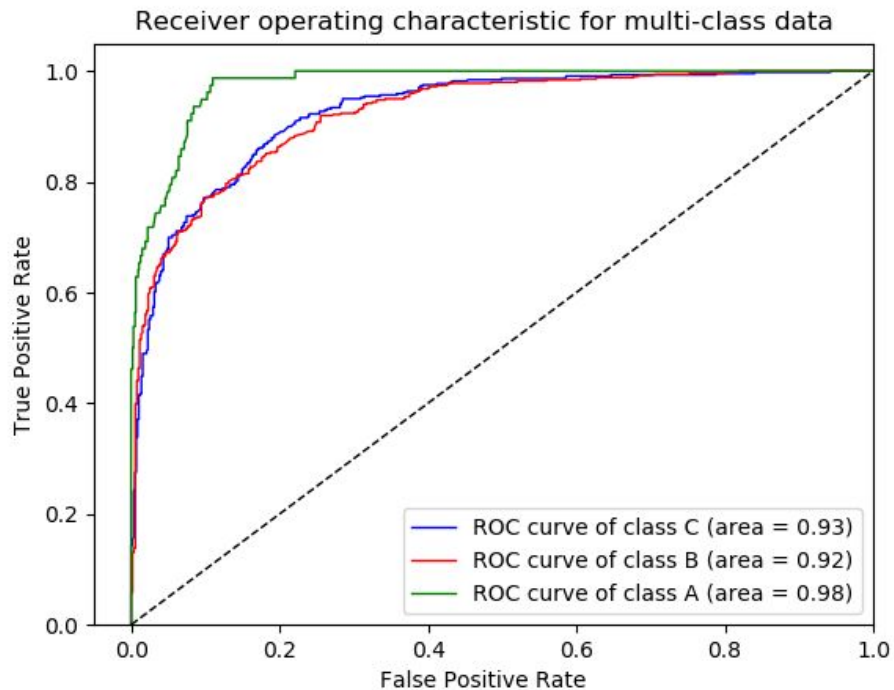
Result

Method	Acc (%)
Gaussian Naive Bayesian	58.5%
Multinomial Naive Bayesian	71.5%
Logistic Regression	82.2%
Random Forest	78.7%
SVM with Linear Kernel	76.9%



Group	Precision	Recall	F1
A	0.86	0.63	0.73
B	0.81	0.84	0.83
C	0.83	0.84	0.84

Result: **ROC/AUC Curve** of Logistic Regression



Result: Linguistic Features

Category (all neg. correlated)	<i>p-value</i> (A-B Comparison)	<i>p-value</i> (B-C Comparison)
Sadness	0.019	0.044
Swearing Words	< 0.001	0.038
Negative Emotion	< 10^{-10}	0.003
Hate	0.004	0.004
Listen	< 10^{-4}	0.016
Hearing	0.002	< 10^{-4}
Fight	< 10^{-4}	0.029
Suffering	< 10^{-5}	0.001

Mann–Whitney U test; only whose *p-value*<0.05 are shown

Individual-level Regression (top 60)

Rank	University Name
1	Princeton University
2	Harvard University
3	Yale University
3	University of Chicago
3	Columbia University
3	Massachusetts Institute of Technology
7	Stanford University
8	University of Pennsylvania
8	Duke University
10	Johns Hopkins University
10	Northwestern University
12	California Institute of Technology
12	Dartmouth College
14	Brown University
14	Vanderbilt University
16	Cornell University
16	Rice University

Avg. Error = 11.15

Abs. Rank Difference	Coverage
≤ 5	25.5%
≤ 10	49.1%
≤ 15	76.8%
≤ 20	90.7%

Individual-level Regression

category	b	positive_emotion	-4.488196
economics	-10.480840	independence	-4.340398
technology	-10.099504	trust	-4.204855
philosophy	-9.904903	prison	-4.075537
science	-9.095995	death	-4.062548
legend	-7.893339	air_travel	-3.897628
terrorism	-7.818766	military	-3.883823
shape_and_size	-7.616722	sadness	-3.645033
writing	-6.441086	gain	-3.578344
healing	-6.296537	affection	-3.577871
poor	-6.196129	art	-3.521534
reading	-4.572166	divine	-3.294159

dispute	5.213781	shopping	6.907196
eating	5.299780	violence	7.174407
fun	5.563269	body	7.253834
crime	5.588866	ocean	7.748175
vehicle	5.716203	noise	8.411105
negative_emotion	5.722408	water	8.711101
movement	5.725064	beach	8.946753
social_media	5.798880	cold	9.078700
animal	6.025168	toy	9.198792
urban	6.396589	ship	9.356454
vacation	6.637485	weather	9.904864
sports	6.842517	play	14.124681

Related Work

- Many tried to analysis & improve university ranking systems, predict based on the previous data, propose new system
 - (Soo, 2005), (Sadlak & Liu, 2007), (Taylor et al, 2007), (Lukman et al, 2010), ...
- Its social impact is **huge**
 - “University ranking as **social exclusion**” (Amsler et al, 2012)
 - “The Dilemmas of Ranking” (Altbach, 2016)
 - “Fatal attraction: Conceptual and methodological problems in the **ranking** of universities by bibliometric methods” (Raan, 2005)
- None of them are text-based

Controversy over Existing Rankings

“Largely managed by non-state organisations in publishing industry or within universities themselves, **ranking has become a form of regulation as powerful in shaping practical university behaviour** as the requirements of States”

- *Simon Marginson,*
professor of international higher education
at the University of London, UK.

Questions?



1	Data Retrieval & Pre-processing	<ul style="list-style-type: none">• Twitter - One of public datasets• Rankings - via crawling
2	LDA with word-vectors	<ul style="list-style-type: none">• LIWC or the like
3	Ordered logistic regression	<ul style="list-style-type: none">• "Ordered" - to reward close estimation
4	Deviance Analysis	<ul style="list-style-type: none">• Explanatory analysis

@jaeseok CS492D
Spring 2019 at **KAIST** is fun! **Presenting** our **analysis** of university rankings this semester with **NLP** and **ML** today.

What are the *linguistic features* involved with *prestigious universities*?

By Jaeseok & Moonyeong