Graduation Project:

Analyze websites' text content using classification

Mef2103

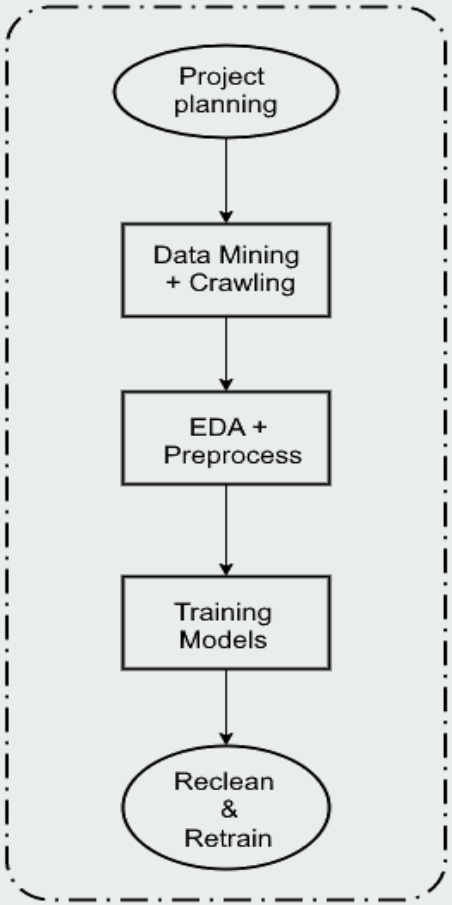
Student : Huynh Truong Tu

About me

“I am Huynh Truong TU, currently a senior student at *HCM city university of Technology* and on-going DIVE INTO CODE course. With the guidance from the instructors, I managed to finish all the sprint assignments and now this is my graduation project. Thank you very much!”

- The overall of this assignment is that it is a NLP-related project where the models will process the text content from a website then predict which industry the website belongs to.
- I use pure statistics of the “key words” of each industry field so that the models can “weight” the class of an sample base on the number of “key words” it holds.
- The output of this project can be used as input for other fields such as marketing or data analytics.

Overview



Project Planning

- Goal: Classify website in predefined list of industry
- Input: Text of a webpage -> Output: Industry Field

Data mining & Crawling

- Find dataset that have domain & it's text & labeling
- Use crawling tools to fill the missing features

EDA + Preprocess

- EDA: use statistic to collect insights of the data corpus
- Using NLP tools to preprocess data from EDA result (remove html tags, stop words, symbol, lemma, ...)

Training models

- Use TF-IDF to find & weight the "keywords" and use them for classification.

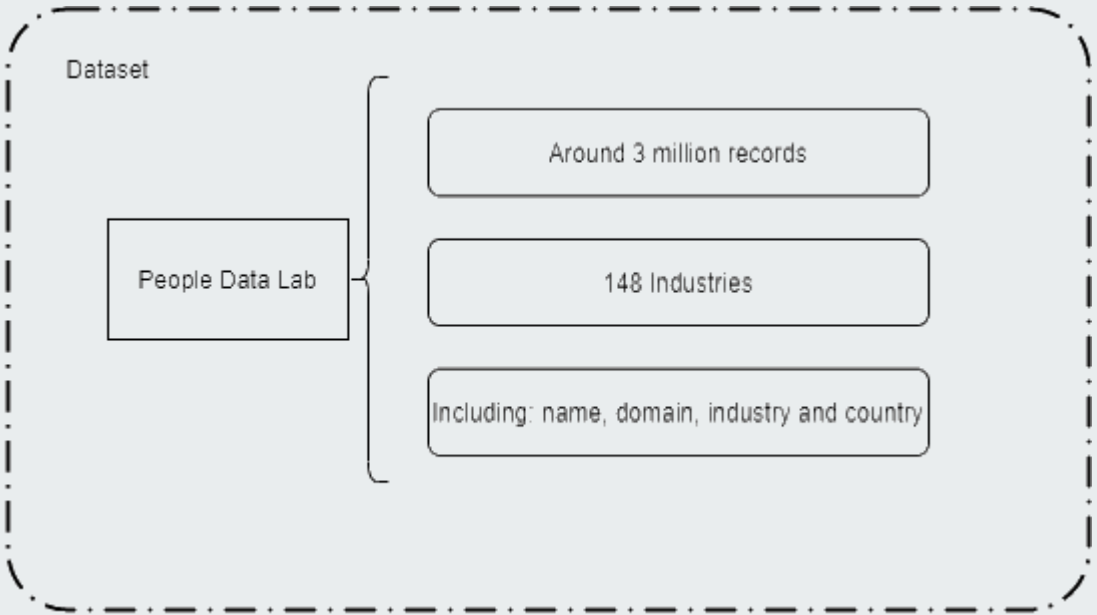
Reclean & Retrain

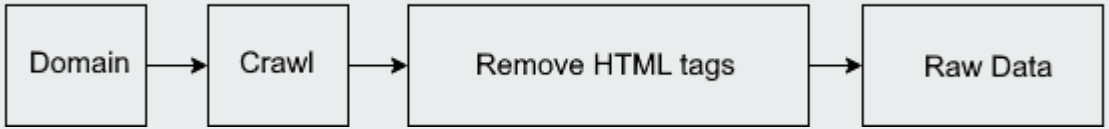
- Base on the results, make adjustments and retrain.

Dataset

- Bought from People Data Lab (a Data company) which includes millions of records

- It doesn't contain text of the domain so crawling text data from domain is required using tools.



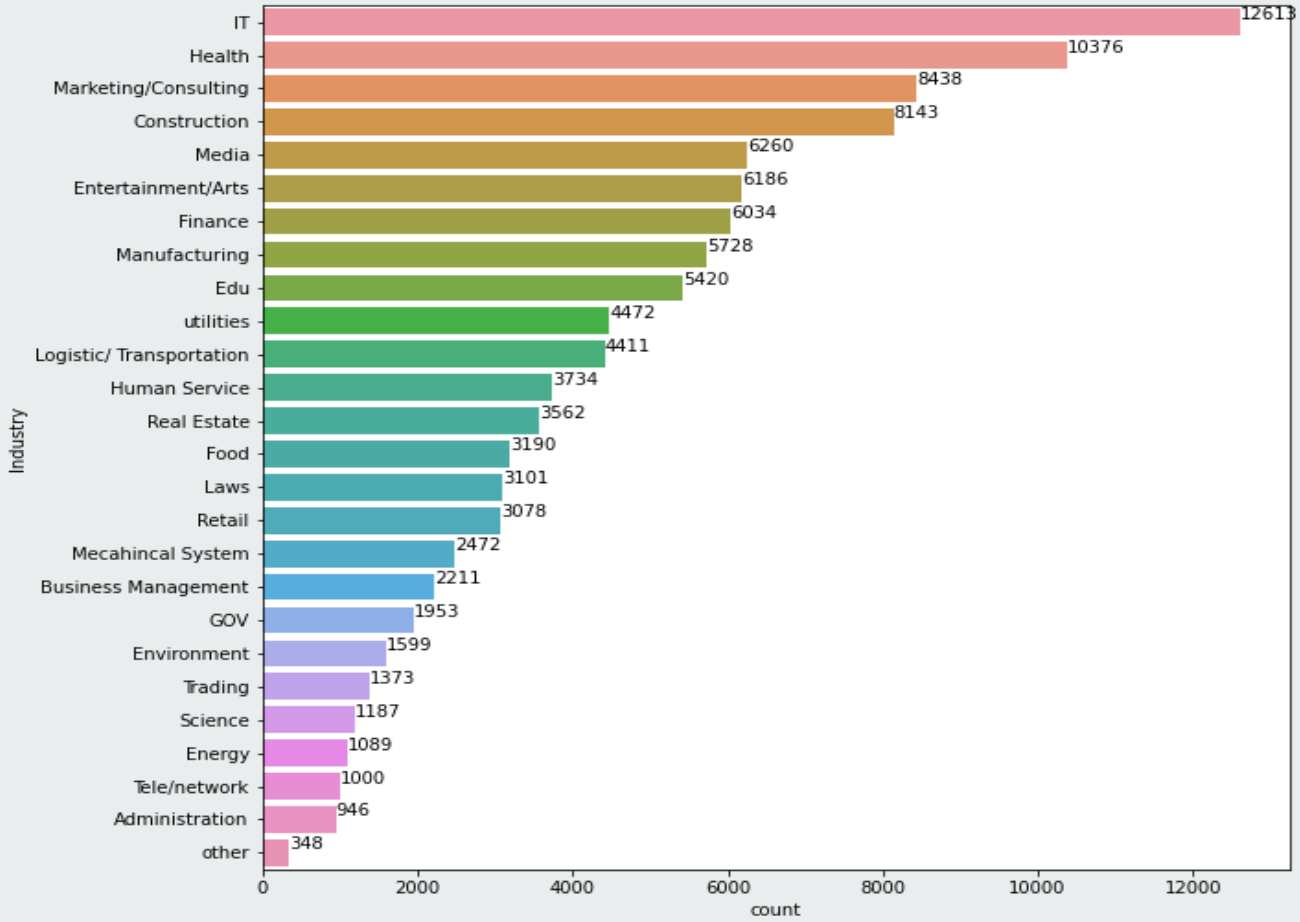


Dataset

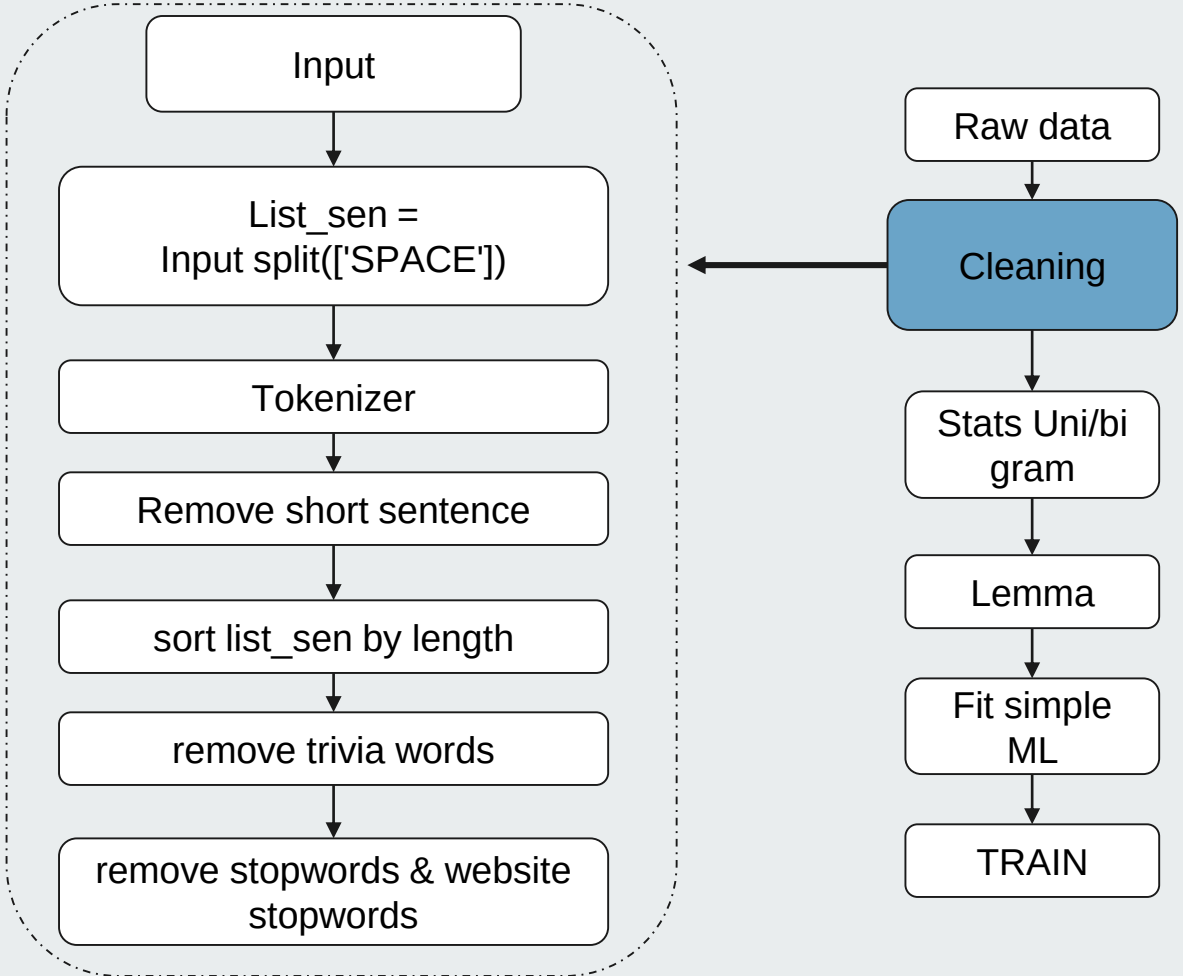
	domain	industry	html
0	balteau.com	mechanical or industrial engineering	Balteau NDT the answer to your X-ray solutions...
1	exnihilo.in	marketing and advertising	This site requires Javascript to work please e...
2	watsontowncmachurch.org	religious institutions	Watsontown Alliance Church wachurch ptd net 57...
3	uwzorgcompaan.eu	hospital & health care	Uw ZorgCompaan 24 uurs aupair zorg aan huis Co...
4	ogcny.com	events services	Order of Good Cheer About Past Events Contact ...
...	...	...	...
30562	orosportsusa.com	sporting goods	Cooling Vests Oro Sports USA Skip to content S...
30563	3dotsdesign.in	marketing and advertising	Why 3 Dots Design is one of the top ad agencie...
30564	scottgaileycpa.com	accounting	Scott Gailey CPA Tax Preparation Services for ...
30565	idealtravelagency.net	airlines/aviation	Ideal Travel Agency Home Expand collapse navig...
30566	aa-planadvies.nl	public policy	AA-Planadvies brengt mensen en idee n bij elka...

Dataset

- 26 main industry
- 148 sub industry
- For this mini-project
  → Predict main industry



Preprocess


Input
List_sen =
Input split(['SPACE'])
Tokenizer
Remove short sentence
sort list_sen by length
remove trivia words
remove stopwords & website stopwords
Raw data
Cleaning
Stats Uni/bi gram
Lemma
Fit simple ML
TRAIN

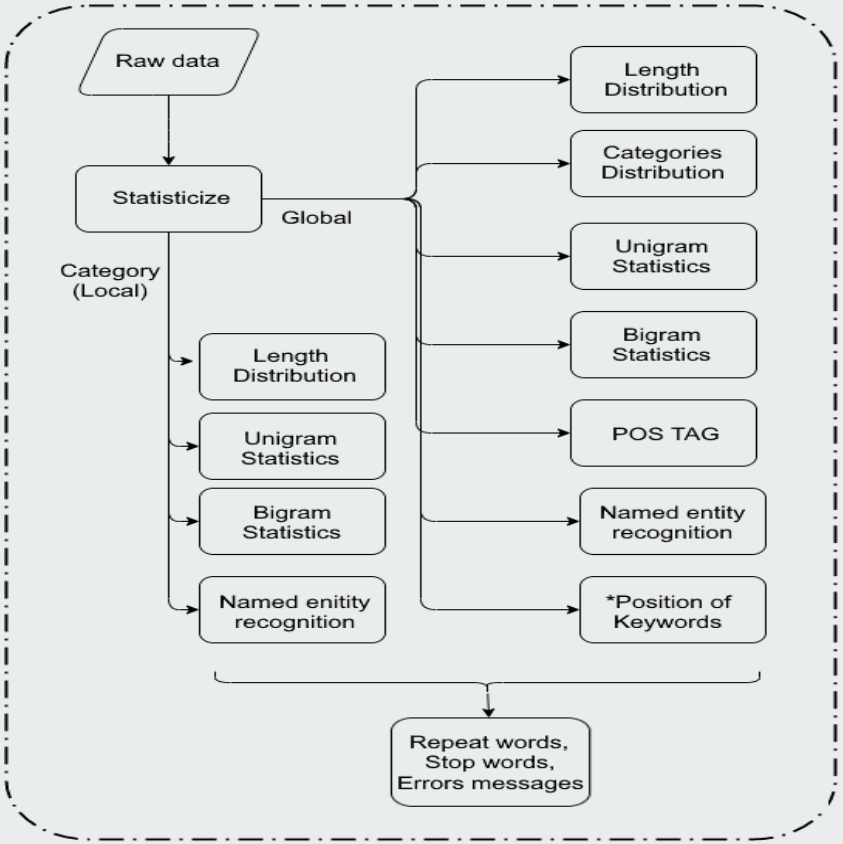
Preprocess

refrigeration control | united states | advance electronic concepts[SPACE]Advance Electronic ConceptsHomeAboutProductsContactWiring DiagramsMoreAdvance[SPACE]Electronic[SPACE]ConceptsSAVINGS. PROBLEM SOLVED.Let's Get StartedWHAT WE DOAt AEC, we design and produce custom-engineered and user-friendly supermarket refrigeration case control systems. Energy efficient, cost-effective, and built to last, our products are developed from the ground up to fit the unique needs of our clients.Learn MoreWHAT MAKES US DIFFERENT[SPACE]Environmentally friendly, economical, durable products[SPACE]Custom solutions available[SPACE]Fast turnaround times[SPACE]70+ years of combined engineering experience[SPACE]Proudly made in the USA[SPACE]OUR QUALIFICATIONSAEC is proud to hold many national and international energy and safety certifications such as:WANT TO LEARN MORE?If you have a question about any of our products or would like to learn more about how we can turn your supermarket headaches into supermarket solutions, contact us today!Contact UsPortland, MEsales@advelecon.com207-797-9825© 2018 by Advance Electronic ConceptsProudly created by Travis Simonds OriginalsHomeAboutProductsContactWiring DiagramsMore

SAMPLE

refrigeration control united state advance electronic concept hold many national international energy safety certification question product turn supermarket headache supermarket solution advance electronic travis problem doat design produce supermarket refrigeration case control system energy efficient built last product ground fit unique make different refrigeration control united state advance electronic concept environmentally friendly economical durable product combined engineering experience

EDA


Raw data
Statisticize
Global
Category (Local)
Length Distribution
Categories Distribution
Unigram Statistics
Bigram Statistics
POS TAG
Named entity recognition
*Position of Keywords
Length Distribution
Unigram Statistics
Bigram Statistics
Named enitity recognition
Repeat words, Stop words, Errors messages

N-gram Language Model

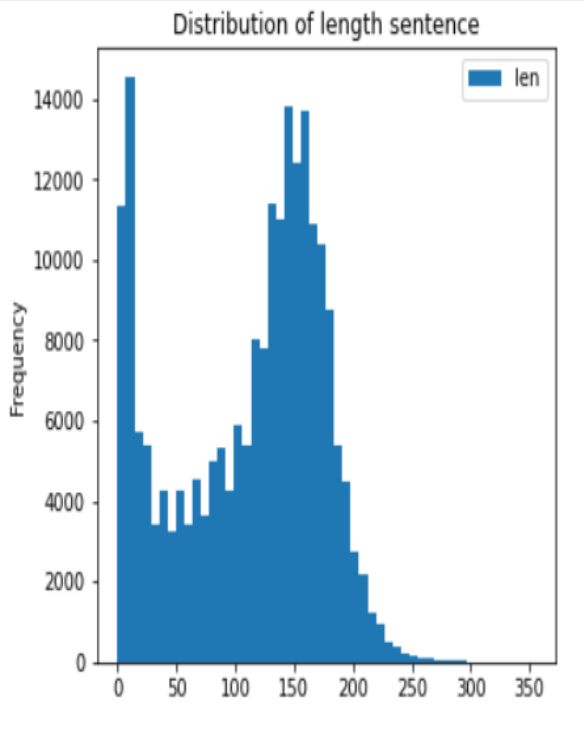
An N-gram is a sequence of N tokens (or words).

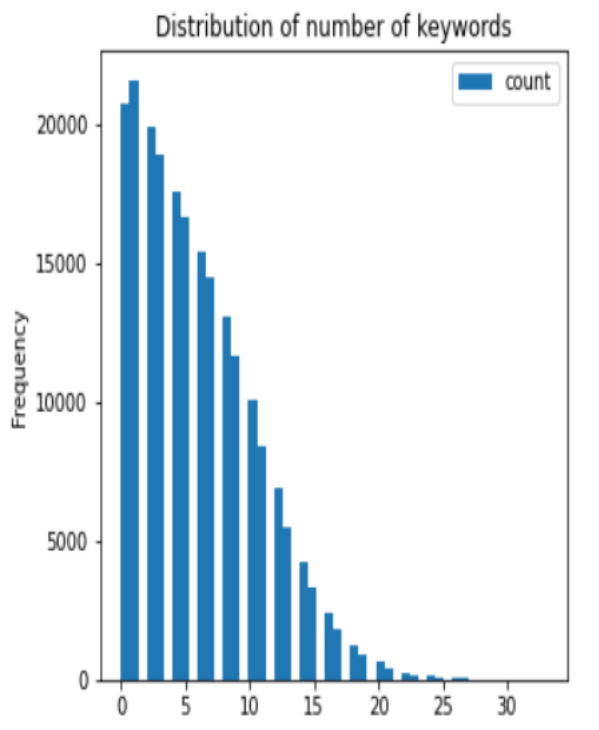
“I learn data mining.”

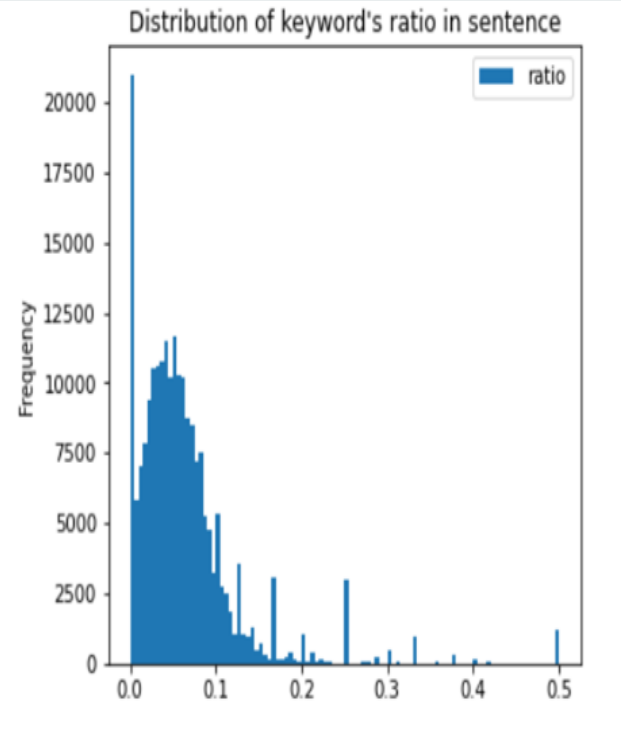
- A 1-gram (or **unigram**) is a one-word sequence:
  “I”, “learn”, “data”, “mining”.
- A 2-gram (or **bigram**) is a two-word sequence of words:
  “I learn”, “learn data”, “data mining”. -> bigger size regarding bags of words

➢ In this mini project concept (predict using occurrence of keywords - TfIDF) → mainly using unigram

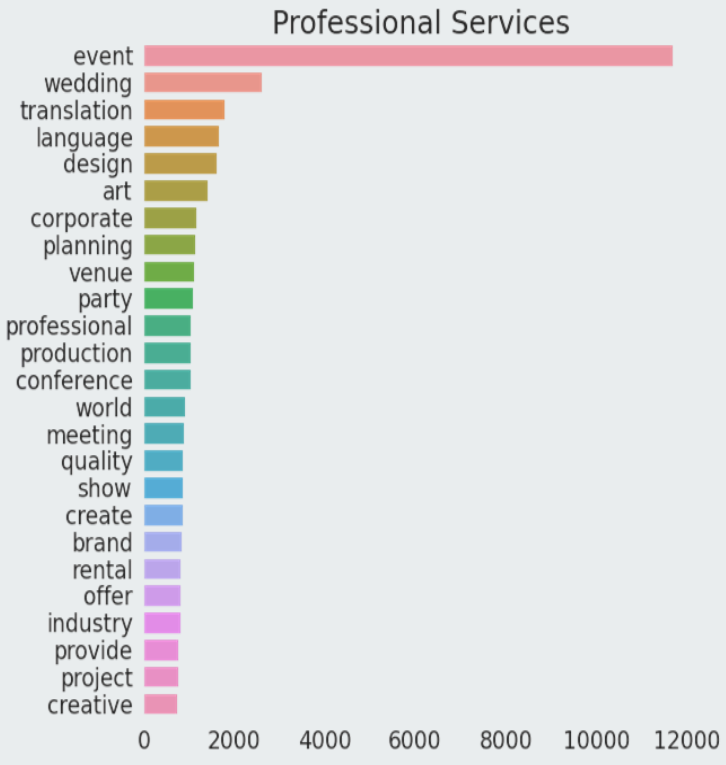
EDA

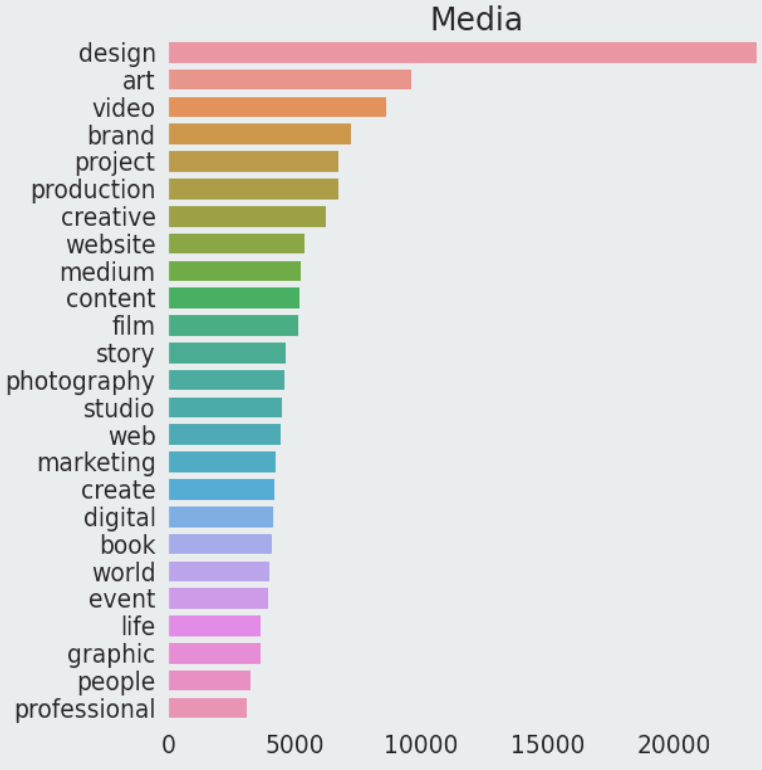

Distribution of length sentence
len
Frequency
0
2000
4000
6000
8000
10000
12000
14000
0
50
100
150
200
250
300
350


Distribution of number of keywords
count
Frequency
0
5000
10000
15000
20000
0
5
10
15
20
25
30


Distribution of keyword's ratio in sentence
ratio
Frequency
0
2500
5000
7500
10000
12500
15000
17500
20000
0.0
0.1
0.2
0.3
0.4
0.5

EDA


Professional Services
event
wedding
translation
language
design
art
corporate
planning
venue
party
professional
production
conference
world
meeting
quality
show
create
brand
rental
offer
industry
provide
project
creative
0
2000
4000
6000
8000
10000
12000


Media
design
art
video
brand
project
production
creative
website
medium
content
film
story
photography
studio
web
marketing
create
digital
book
world
event
life
graphic
people
professional
0
5000
10000
15000
20000

Validation Results

- Uni TF-IDF
  - Linear Support Vector (loss : 'squared_hinge')
    - Acc = 0.6666666666666666
  - Stochastic Gradient Descent (loss function: 'squared_hinge')
    - Acc = 0.7569504310344828
  - Logistic regression (loss: ‘Log Loss’)
    - Train Acc = 0.841
    - Validate Acc = 0.734
  - Voting Classifier (LR, SGDC)
    - Acc = 0.7863673660766616
- Bi TF-IDF
  - SGD Classifier (loss function: 'squared_hinge')
  - Acc = 0.679138500235812

Extract Top Keywords

Business Services: human, role, placement, assistant, people, career, hiring, human resource, employee, employer, outsourcing, recruiter, executive, talent, recruiting, job, staffing, hr, candidate, recruitment

Construction/Architecture: renovation, flooring, stone, civil, structural, roof, plumbing, door, installation, architectural, furniture, project, builder, roofing, building, architecture, contractor, concrete, architect, construction

Consulting: thinking, coaching, organizational, value, success, process, advisory, innovation, implementation, improvement, growth, organisation, consultancy, strategic, leadership, organization, change, consultant, strategy, consulting

Consumer services: pilate, chiropractic, supplement, skin, exercise, gym, animal, workout, health, nutrition, beauty, hair, veterinary, massage, wellness, body, salon, pet, yoga, fitness

Education: teacher, educational, campus, parent, language, university, academic, admission, career, coaching, child, training, library, skill, college, course, education, student, learning, school

Entertainment/Arts: recording, league, season, dj, gallery, football, show, ticket, club, art, soccer, game, play, sport, entertainment, player, theatre, dance, artist, music

Environment/Energy: soil, environment, oilfield, pipeline, carbon, conservation, renewable, sustainable, tree, drilling, water, recycling, fuel, oil gas, gas, oil, waste, solar, energy, environmental

Finance/Trading: risk, benefit, money, mortgage, advisor, trading, investing, payment, retirement, bank, finance, banking, credit, fund, loan, wealth, investment, capital, financial, insurance

Food & Baverages: organic, bakery, cheese, cake, menu, chocolate, delicious, meat, order, flavor, fruit, ingredient, farm, fresh, taste, beer, coffee, restaurant, wine, food

⋮

Limitations & Future developments

- Limitations:
  - TF-IDF: purely statistics - occurrences of keywords.
    → Newspaper domain: low accuracy
    → Can't make use of pos tag & named entities
    (after stemming & lemmatization the structure is lost)
    → Some industries with relative fields may be confused
  - The scope is limited in English domains.
  - Output depends on the categories of industries
    (Each country has different way of industrial classification.)
- Future development:
  - Combine with CVision.
  - Replace TF-IDF with BERT from Google
    (pre-train model, highly accurate, applicable for many languages including Vietnamese)

THANK YOU FOR LISTENING