Capstone Project

Company Classification

Content

- Problem statement
- Data summary
- Covering insights
- Challenges
- Conclusions

Problem Statement

We are given with web scraped data of various businesses and companies.

We need to somehow categorize these businesses and companies across a standard taxonomy (consists of term names and labels that are specific to an organization's information and unique to how that business operates). So that, business can leverage this information and target potential companies.

Data Summary

- We have records of 73974 companies with 9 columns/features.

- Each feature is a text data.

- We will be left with only 4K records, if we remove all the null values at the first place.

Data Overview

- **Website:** The website of the company/business
- **Company Name:** The company/business name
- **Homepage Text :** Visible homepage text
- **H1:** The heading 1 tags from the html of the home page
- **H2:** The heading 2 tags from the html of the home page
- **H3:** The heading 3 tags from the html of the home page
- **Navlink text:** The visible titles of navigation links on the homepage (Ex: Home, Services, Product, About Us, Contact Us)
- **Meta keywords:** The meta keywords in the header of the page html for SEO
- **Meta description:** The meta description in the header of the page html for SEO

Techniques Implemented

In the given data there were more null values equally spread across all the major columns. To avoid data loss we have combined all the columns as we are dealing with text.
We also analysed individual columns to gain insights.

Data Preprocessing

- Handled null values
- Removed punctuations
- Removed digits
- Removed stop words
- Removed special characters
- Stemming/Lemmatization
- Removed short words

Word Embedding

Word embedding is a term used for the representation of words for text analysis

Few popular techniques are:

- Countvectorizer
- TF-IDF Vectorizer
- Word2Vec

We experimented with different techniques and out of which TF-IDF gave us better results.

Using K Means

- Kmeans is an unsupervised machine learning algorithm used for classification.

- To solve this problem statement, we went ahead with K Means algorithm to categorize the companies into multiple clusters.

Clusters Insights

INSIGHTS FROM OVERALL PERSPECTIVE..

Optimal number of clusters..

By using Elbow method and Silhouette score, we came up with 12 as an optimal of clusters.

Elbow method

Cluster 1 : PET HEALTHCARE

Example sentences

Company Name	Text
Pleasant Valley Animal Hospital	adopting a pet bereavement loss of a pet..cancer chemotherapy cardiology cat only..
Annabessacook Veterinary Clinic	home meet our team large Animal Services Small Animal Services KDM Farm Resources & Externships..
Columbia Pike Animal Hospital	Request an Appointment Prevent Animal Cruelty Shop Online Access Pet Records..
Doylestown Animal Medical Clinic	proud to provide a full spectrum of routine and specialized medical services designed to care for pets..
Borchard Veterinary Clinic	PET WELLNESS & PREVENTATIVE CARE PET SURGERIES..

Cluster 2 : IT SERVICES

Example sentences

Company Name	Text
Center for Computer resources	Services & Solutions Microsoft Azure Microsoft Teams Modern Desktop Microsoft 365 Business..IT Consulting Webroot showcase Virtualization Projects..
Appmetric software limited	We believe that software consulting is more of a tactic than a company..reckoning entity in the competitive field of software services and business process outsourcing..
Fortune business transfers & acquisitions	Services Business Acquisition Consultants Preparing to Sell Business Help Selling Business..
Construe solutions	Wireless Solutions Services & Consulting Business consultant API Integrations Mobile Application..

Cluster 3 : INSURANCE

Example sentences

Company Name	Text
Stratford insurance group	We provide personalized support and full-service insurance to individuals and businesses to ensure you are protected in all your adventures..
Petley-hare limited insurance brokers	You're Unique. Your Insurance Should Be Too!
Private client insurance group	We are a global specialty insurance distributor with expertise in property, casualty, professional lines and specialty group benefit products..
W.C. Burgess insurance services ltd	For all your insurance needs, whether it is for your car, home, boat, snowmobile, ATV, business, life or disability insurance we have the answers you need and the prices you want..

Cluster 4 : DIGITAL MARKETING

Example sentences

Company Name	Text
The Digital Marketing People Inc.	Marketing is much more than just a great website..Work with a Digital Marketing Expert to use the right mix of digital marketing services..
ITI marketing	We provide analytical reports that permit clients to better tailor their off-line corporate advertising and marketing strategies to those demographics most responsive to their products..
Allegra Pittsburgh Design & Marketing	We're a full-service marketing and print communications company based in Pittsburgh, PA..
Renegade Advertising Inc	Renegade Advertising is a full service internet advertising agency..Renegade Advertising focuses on online and new media marketing and advertising..

Cluster 5 : LOGISTICS

Example sentences

Company Name	Text
Alpha Global Transport Limited	runs a tight ship with Transport Captain Keith Bonner..experience in both the air freight and general haulage industry..
Clark logistic services	Our flexible warehousing solutions and distribution services are focused on delivering a complete turn-key solution for your business-to-business order fulfillment needs..
Auction Transport services inc.	Leaders in Vehicle Transportation..
Eal & aw Transport group	our fleet consists of over 30 specially equipped trucks and specially equipped trailers to cater for the delivery of windows and doors in both the domestic and commercial market..

Cluster 6 : HEALTHCARE (EYE CARE)

read
learn
eye care
clinic
health care
medic center
hospit
provid
practic
treatment
servic
program
welcom
primari care
urgent care
product
help
well
doctor
work
new
team
consult
special
privaci polici
patient portal
medic group
partner
see
experi
physician
support
care servic
inform
solut
offic
locat
eye exam
find doctor
year
visit
includ
staff
train
need
home
care patient
facil
right
reserv
know
use
pay bill
physic therapi
start
share
system
review
manag
mani
dri eye
pharmacy
new patient
patient care
boss appoint
appoint
benefit
home care
time
line
procedur
result
sinc
healthcar busi
make appoint
schedul appoint
dr
client
dr eye
trust
understand
advanc
complet
one
area
pa pa
check
weight loss
chos
develop
compani
design
text
specialist
focus
avail
accessibl
community
time
dr
first
time
time
time

Cluster 7 : FINANCIALS

Example sentences

Company Name	Text
Weitzel Financial Services Inc	(WFS) has been to help our clients reach their financial goals by providing the support and resources needed to make sound, individualized financial decisions..
Greenlink Financial	Leader in consumer unsecured personal loans, helping thousands of people resolve their consumer debt with a personal loan..We've helped thousands of people resolve their debt and find financial freedom. Start Living Your Life Today!
Northwest financial & tax solutions, inc.	we offer a complimentary second opinion and comprehensive portfolio analysis..
Trinity tax service	We offer a broad range of services for business owners, executives and independent professionals..Our services encompass nearly every aspect of financial life.

Cluster 8 : REAL ESTATE & LEGAL AFFAIRS

Example sentences

Company Name	Text
Morris Law group	Morris Law Group is a boutique estate planning and asset protection law.. Through effective planning, our experienced attorneys are dedicated to helping individuals and families preserve and protect their wealth for future generations..
Premier (uk) property management ltd.	portfolio of listed and off-market properties ranging from exclusive domestic and commercial premises to hotels, land and development projects..
White cleland lawyers & consultants	Our Legal team is already highly skilled, but are each equally committed to ongoing practical learning..When a relationship breaks down you need support and assistance in working through the myriad of potential legal issues
Blackstone williams properties, llc	Blackstone Williams provides first-class rentals, sales & property management in Boston's Back Bay & beyond!

Cluster 9 : HOSPITALITY

Example sentences

Company Name	Text
hawaiian hotels & resorts	breathtaking beaches and incredible natural wonders. Explore Now the best of Hawaii Escape to The Big Island for an unforgettable vacation or romantic honeymoon in the heart of
refinery hotel	imaginative reinterpretation of the Colony Arcade,Äîa former hat factory and Prohibition-era tea room. More than 100 years of Garment District history are on display at our luxury hotel near
the wellington apartment hotel,	Accommodation in Kangaroo Point near the Gabba, East Brisbane.¬†Perfect Self Contained Apartment Accommodation for holiday and corporate travellers, groups and conferences near Brisbane CBD. Come and experience our Luxury Central Brisbane
lower plenty hotel,	Award-winning and welcoming Friends, family or work colleagues. No matter who you,Äôre dining with, our Bistro is the perfect place to entertain your tastebuds. Named the Best

Cluster 10 : E-COMMERCE

Example sentences

Company Name	Text
PrestaShop	freemium, open source e-commerce platform..You can launch your online store right now and start to sell online..
WooCommerce Pvt ltd	e-commerce plugin for WordPress. Designed for small to large-sized online merchants using WordPress..
OXID eShop	ecommerce solution built using object oriented programming and PHP..
Currys	British electrical retailer operating in the United Kingdom..Free Delivery or Order & Collect..

Cluster 11 : HOME SERVICES

Example sentences

Company Name	Text
auction transport services inc.	HOME ABOUT SERVICES CONTACT EVENTS JOBS FranVßais AUCTION TRANSPORT SERVICES,customer service driven organization, where safety and consistently clean deliveries¬†are always a priority,
custom electric sd	Custom Electric SD. We handle all of your electrical contracting needs. We specialize in recessed lighting, and both residential and commercial wiring
d-tech electrical contractors ltd	by our NICEIC Approved Contractor, Construction Line and CHAS accreditations. All of our directly employed electricians are full qualified and highly experienced, which enable us to maintain our
electrical installations inc,	is what makes Electrical Installations shine. Our expertise as an electrical contractor adds an additional dimension of experience and capability that other system integrators cannot offer. Our

Cluster 12 : RENEWABLE ENERGY & UTILITIES

Example sentences

Company Name	Text
Commercial & Industrial Microturbine,	Industrial-Scale Alternative Energy Developer & Financier,reduce their energy supply costs, and¬†become more sustainable and resilient
solarpowergb.com	Welcome to Solar Power GB, the premier provider of solar PV, solar thermal, solar panel cleaning and energy saving products in Essex. ",Renewable Energy
appleblossomenergy.com	Insulation Duct Insulation Reflective Barriers Wall Insulation Floor Insulation Insulation for Builders Cellulose Insulation
sd-windenergy.com	No Shut Down Wind Turbines#sep#Share this page#sep#News & Updates#sep#A Unique Delta Rotor Design",Renewable Energy

Distribution of Categories

INSIGHTS FROM HOMEPAGE..

Optimal Clusters

By using Elbow method and Silhouette score, we came up with 9 as an optimal number of clusters.

Cluster 1 : ENERGY & UTILITIES

Cluster 0

Cluster 2 : INFORMATION TECHNOLOGY

Cluster 1

Cluster 3 : FOOD DELIVERY SERVICES

Cluster 4 : REAL ESTATE & LEGAL

Cluster 5 : INSURANCE

Cluster 6 : EVENT MANAGEMENT

Cluster 5

Cluster 7 : HEALTHCARE

Cluster 6

Cluster 8 : DIGITAL MARKETING

Cluster 7

Cluster 9 : HOME SERVICES

Cluster 8

Distribution based on ‘Homepage_txt’ Column

INSIGHTS FROM HEADERS..

Optimal Clusters

By using Elbow method and Silhouette score, we came up with 12 as an optimal number of clusters.

Cluster 1 : REAL ESTATE

Cluster 2 : LEGAL AFFAIRS

Cluster 3 : FOOD SERVICES

Cluster 4 : IT SERVICES

Cluster 5 : LOGISTICS

Cluster 6 : INSURANCES

Cluster 7 : HEALTHCARE

Cluster 8 : INDUSTRIALS

Cluster 9 : EATERY INDUSTRY

Cluster 10 : DIGITAL MARKETING

Cluster 11 : HOSPITALITY

Cluster 12 : ENERGY & UTILITIES

INSIGHTS FROM
Meta_keywords & Meta_description..

Optimal Clusters

By using Elbow method and Silhouette score, we came up with 9 as an optimal number of clusters

Cluster 1 : REAL ESTATE & LEGAL AFFAIRS

Cluster 2 : FOOD SERVICES

Cluster 1

Cluster 3 : DIGITAL MARKETING

Cluster 2

Cluster 4 : FINANCIALS

Cluster 3

Cluster 5 : ENERGY & UTILITIES

Cluster 4

Cluster 6 : INSURANCES

Cluster 5

Cluster 7 : HOME SERVICES

Cluster 6

Cluster 8 : INDUSTRIALS

Cluster 7

Cluster 9 : HEALTHCARE

Challenges

- Text Preprocessing as there were characters from multiple scripts.
- There were lot of null values that we had to handle carefully.
- Limited resources : Dataset turned to be large for computation power we have.
- Deciding the number of optimal clusters was bit difficult since the dataset is really vast and can have more number of clusters.

Conclusion

- As we observed “E-commerce” was the top cluster from overall perspective, we can further dig down and do sub-clustering to gain more insights.

- We were able to categorize businesses/companies in 9 to 12 clusters based on different features and on the combination of whole features as well.

- Those clusters represent various industries and we can approach those companies of each cluster with relevant business proposals.

Q & A