I'm unemployed

Goal

The goal of my project is to extract and identify the underlying themes or topics that frequently occur in job descriptions across different categories.

Huh? Why?

Understand current market trends, prioritize learning efforts and portfolio building, align profiles with the topics and skills in demand, position yourself in the job market, discover niches.

How?

By using Natural Language Processing. Mainly, Topic Modeling.

Topic modeling?

A type of unsupervised learning method used to discover abstract topics that occur in a collection of documents.

What is a topic?

In the context of topic modeling, a topic is a collection of dominant, representative keywords.

A document consists of a collection of topics

Topics should be clear, segregated, and meaningful.

We must keep in mind:

The quality of the text pre-processing

The number of topics we want to uncover

The choice of model used and choice of pre-processing

Tuning hyper parameters within our models

But first.

The data.

AHMED MYALO · UPDATED 3 MONTHS AGO

Upwork freelance jobs (+60k)

+60K various jobs scrapped from Upwork site (this year 2023)

RAW Dataset

Job Title	Description	Category_1	Category_2	…	APPLICANTS	Country
Lookin for a data genius	I need to build an AGI before the end of the year	DATA ANALYSIS	NaN	…	40,000	Canada
So much data!	I need someone to big data my data	DATA ANALYSIS	ENTRY LEVEL	…	10,1337	Peru

Filtered Dataset

Job Title	Description	Category_1	Expertise Level
Need Data Scientist Urgently!!!	You must know SQL, PANDAS, PYTORCH, NUMPY, SCIPY, APPLEPIE, EXCEL	DATA ANALYSIS	EXPERT
Need to make an AI pls	excel?	DATA ANALYSIS	ENTRY LEVEL

Pre-processing

Tokenization

Removing stop words

- BERT suggests against this

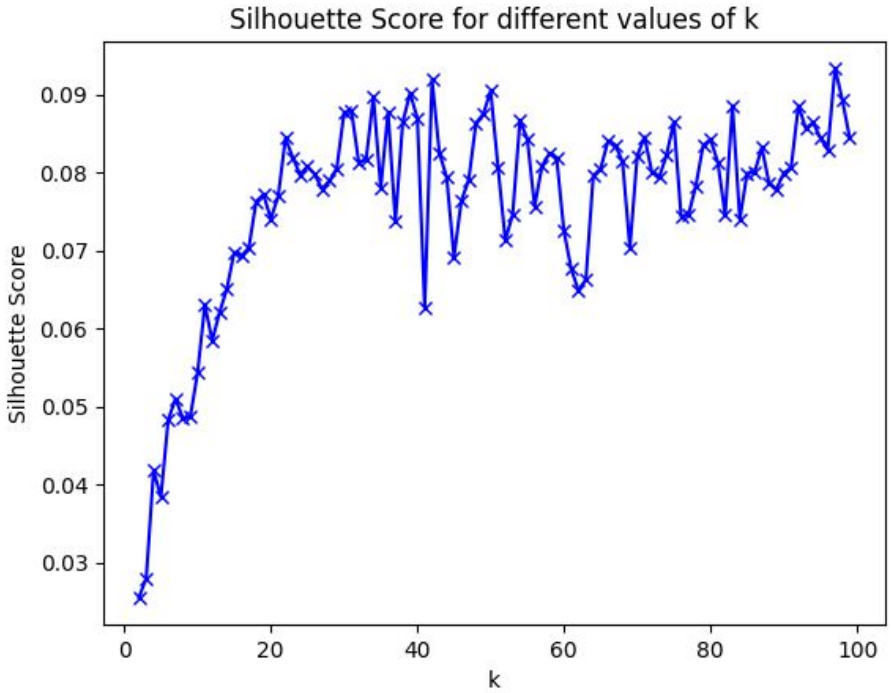
Lemmatization

- When computing LDA

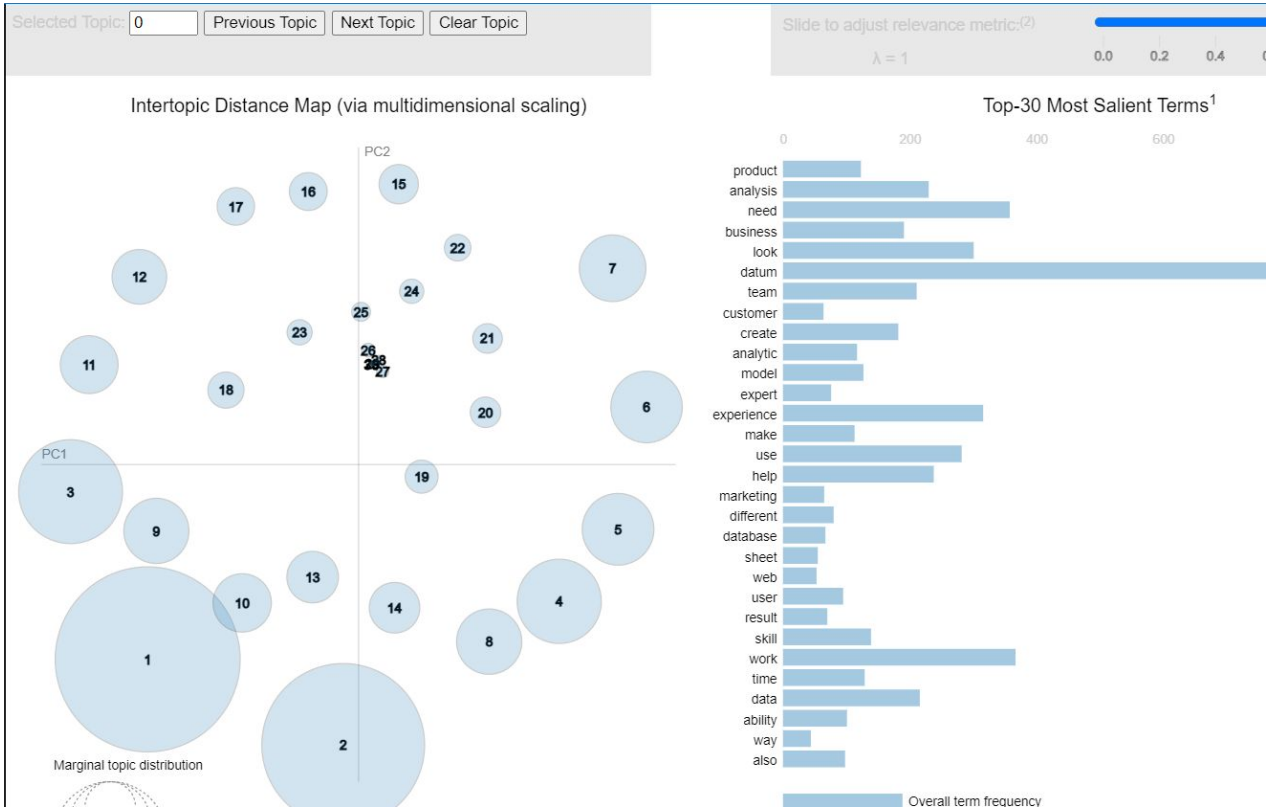
Vectorization

Model 1: TF-IDF + KMEANS

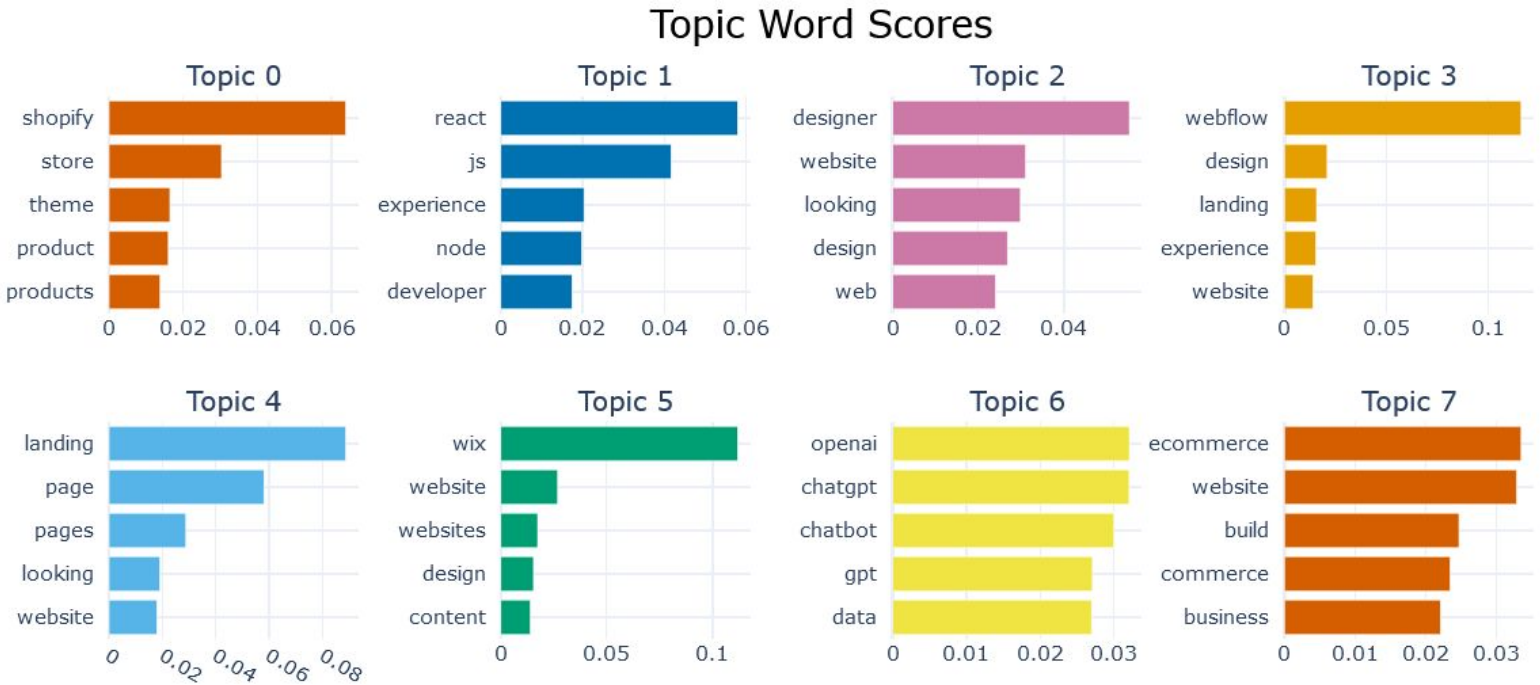
Cluster 0	Cluster 1	Cluster 2
need	google	looking
data	analytics	time
help	product	help
excel	data	work
analysis	tracking	model
project	marketing	like
create	looking	data
job	help	want
code	team	project
report	want	use



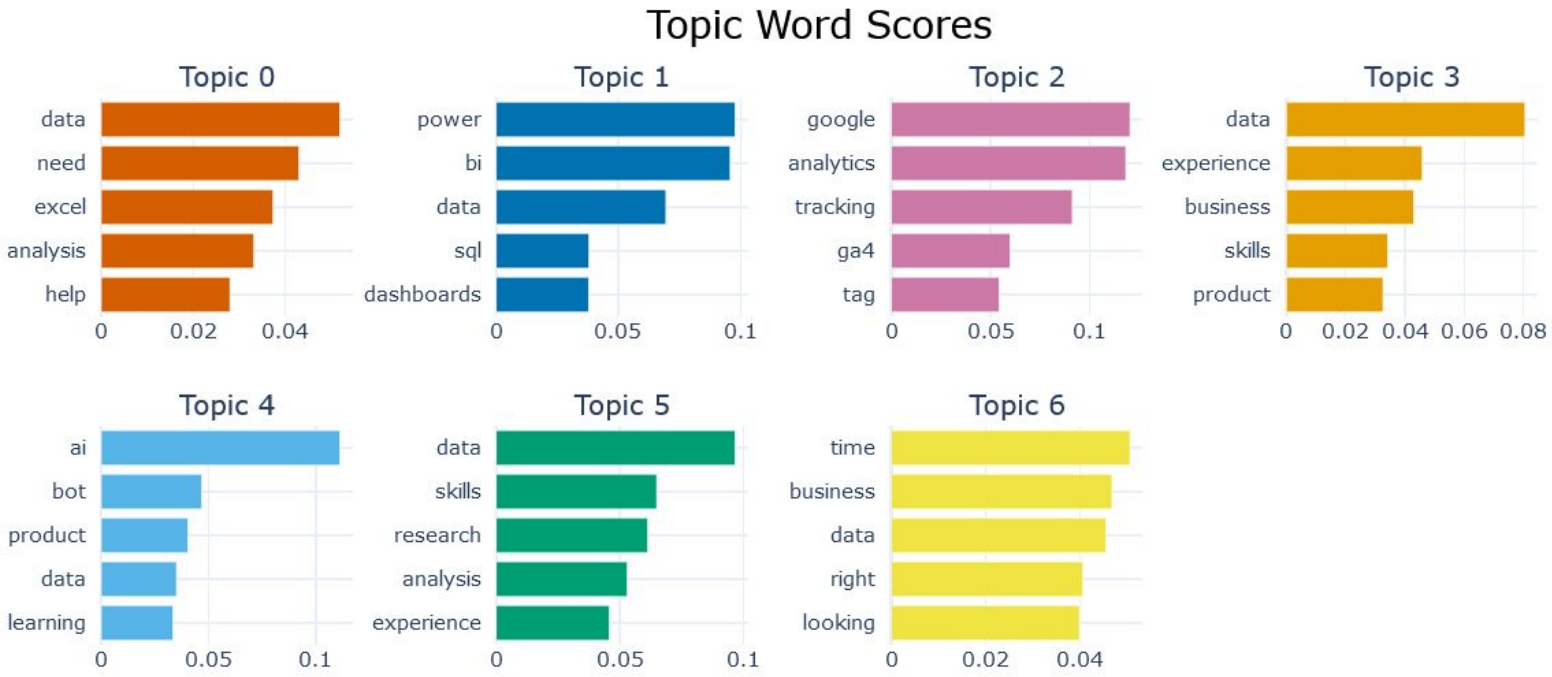
Model 2: LDA


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
Marginal topic distribution
Top-30 Most Salient Terms1
0
200
400
600
product
analysis
need
business
look
datum
team
customer
create
analytic
model
expert
experience
make
use
help
marketing
different
database
sheet
web
user
result
skill
work
time
data
ability
way
also
Overall term frequency

Model 3: BERTopic



BERTopic for Data



Next steps

Expand to other services - Fiverr

See how clusters evolve over time

See how clusters change by geography

Thank you