Searching LinkedIn using Topic Modeling

Team of Moumi Das,Rajarshi Chowdhury,Shiladitya Swarnakar, Soumya Banerjee,Souvick Datta and Sowmya Kartik

A presentation by Group 4
Data Science Batch of July'19

Our agenda

- Introduction
- Customer Pain Point
- Business Problem
- Data Collection
- Data preprocessing
- Data Modelling
- Deployment
- Product Marketing
- Conclusion

0 Introduction

Our moto:
Just Act

Our goal:
Aiming The 'Wow'

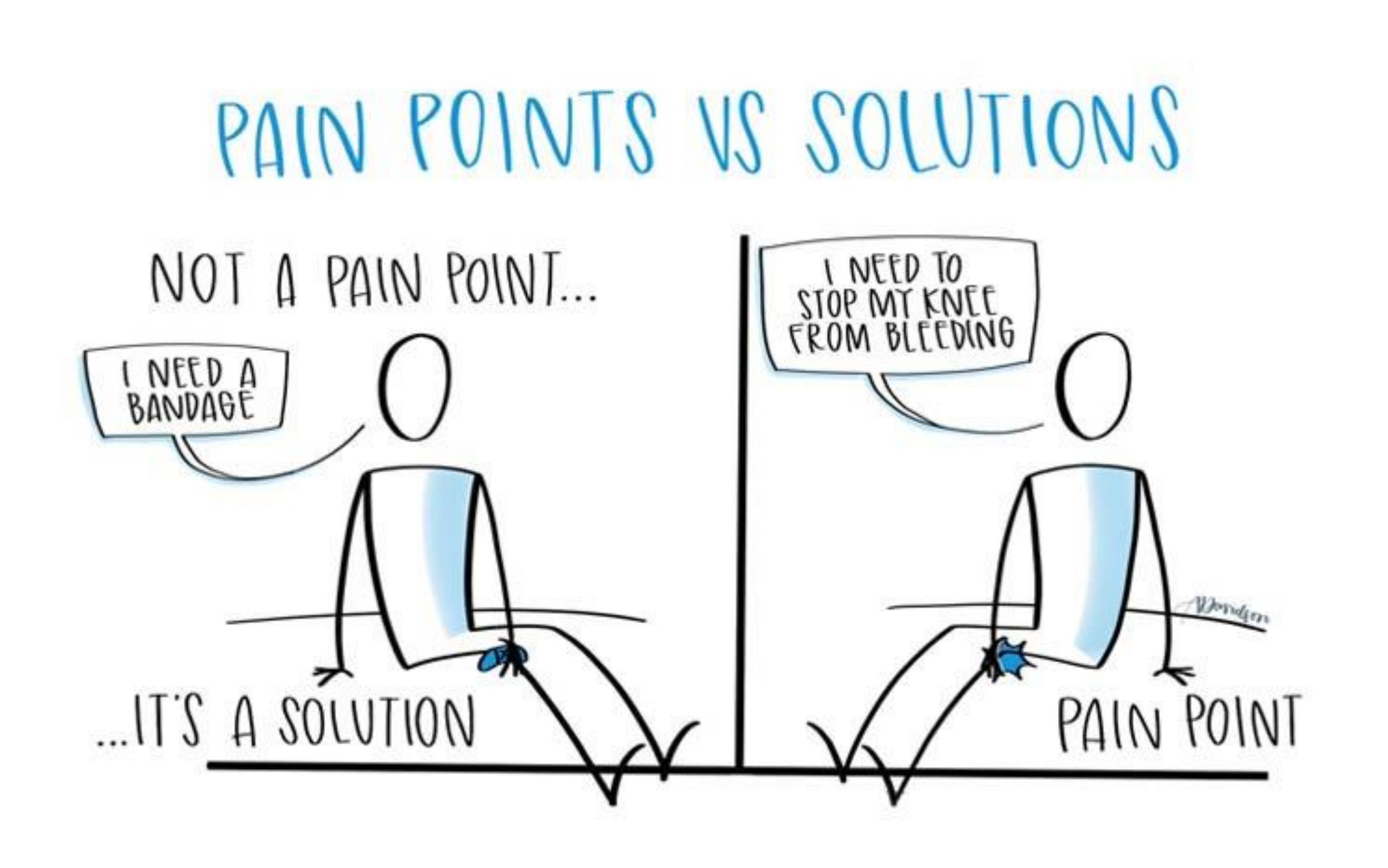
“Content is King” -Bill Gates

Our 3 months in

2 min.

Let's do the engineering problem

"Once confined to fantasy and science fiction time travel is simply an engineering problem" - Michio Kaku


PAIN POINTS VS SOLUTIONS
NOT A PAIN POINT...
I NEED A BANDAGE
...IT'S A SOLUTION
I NEED TO STOP MY KNEE FROM BLEEDING
PAIN POINT

"Why should I buy your product?"

“My friend told me your product is not good

“What’s so different about you?”

"Why not that product?"

“Why do we need you?”

1 Customer Pain point

That product costs less!

“Why should we trust you?”

"I have been using that for so long.."

Reducing asymmetry of information between candidate pool and recruiter pool.

2

Business Problem

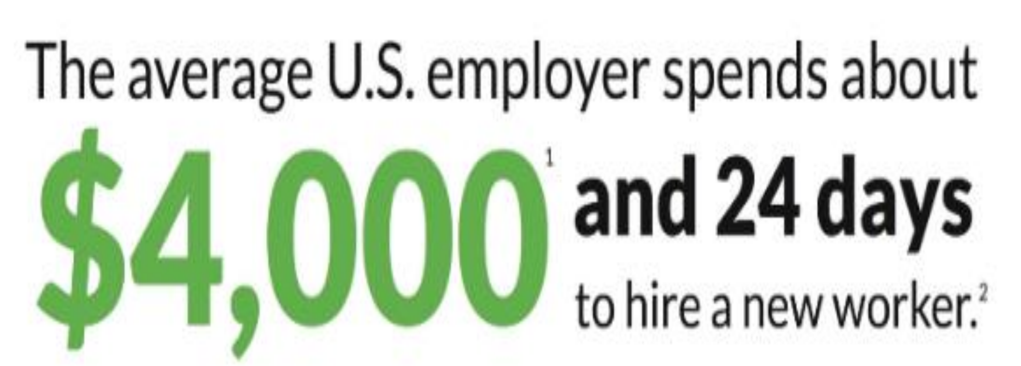
Average cost/hire is

4,129$

42 days

To refill a position

Source: According to Human Capital Benchmarking Report by Society for Human Resource Management



The average U.S. employer spends about

$4,000[1] **and 24 days**

to hire a new worker.[2]

Source: As per GlassDoor , July 5,2019

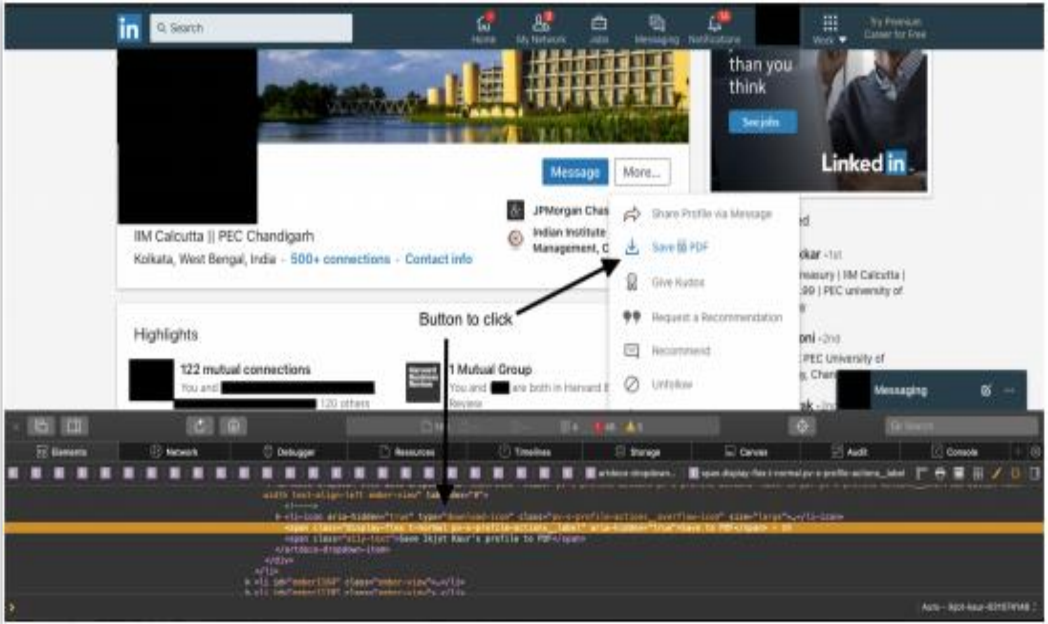
Understanding the problem

- Right resource at the right position can do miracles in business giving
  - Competitive advantage over other players.
- Cost of refilling a position in terms of money and time can be really expensive.
- We try to minimize *search time at a low cost* both for :
  - A recruiter looking for candidate.
  - A candidate looking for opportunities.

2 Data Collection

Data Collection?

Web Scraping



[redacted]

IIM Calcutta || PEC Chandigarh

Kolkata, West Bengal, India · 500+ connections · Contact info

JPMorgan Chas

Indian Institute Management, C

- Share Profile via Message
- Save to PDF
- Give Kudos
- Request a Recommendation
- Recommend
- Unfollow

Button to click

Highlights

[redacted] 122 mutual connections

You and [redacted] 120 others

1 Mutual Group

You and [redacted] are both in Harvard Business Review

```
width text-align-left ember-view" tabindex="0">
    <!---->
  ▶<li-icon aria-hidden="true" type="download-icon" class="pv-s-profile-actions__overflow-icon" size="large">…</li-icon>
    <span class="display-flex t-normal pv-s-profile-actions__label" aria-hidden="true">Save to PDF</span> = $0
    <span class="a11y-text">Save Ikjot Kaur's profile to PDF</span>
  </artdeco-dropdown-item>
 </div>
</li>
▶<li id="ember1304" class="ember-view">…</li>
```

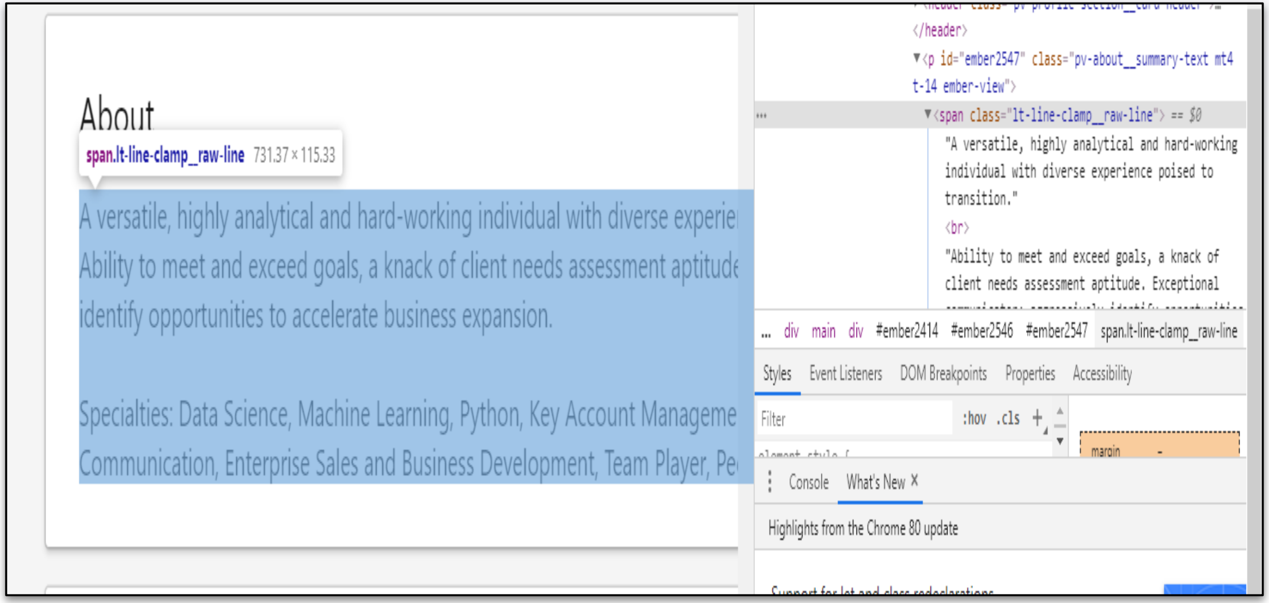
- Extract large amounts of data from websites within a small interval of time.
- Data stored in Python Dataframe for analysis

Challenges faced while data collection

I THINK WE NEED A NEW WAY TO CAPTURE INSIGHTS FROM OUR DATA.

Overcoming the challenge

- Scrape data directly from the LinkedIn.
- Use inspect element to locate the relevant information.
- Approx 5800 public profiles scraped.
- Data stored in dataframe



About

span.lt-line-clamp__raw-line 731.37 × 115.33

A versatile, highly analytical and hard-working individual with diverse experie

Ability to meet and exceed goals, a knack of client needs assessment aptitude

identify opportunities to accelerate business expansion.

Specialties: Data Science, Machine Learning, Python, Key Account Manageme

Communication, Enterprise Sales and Business Development, Team Player, Pe

```
</header>
▼<p id="ember2547" class="pv-about__summary-text mt4
t-14 ember-view">
  ▼<span class="lt-line-clamp__raw-line"> == $0
    "A versatile, highly analytical and hard-working
    individual with diverse experience poised to
    transition."
    <br>
    "Ability to meet and exceed goals, a knack of
    client needs assessment aptitude. Exceptional
```

Highlights from the Chrome 80 update

Tools Used

01 SELENIUM

- Python package used for automated Scraping
- Extracts raw web data

02 BEAUTIFUL SOUP

- Used to refine the text
- Used to filter out the HTML tags from raw text

All with Python API

Data

- Name
- Summary
- Skills
- Experience_1
- Current_Organisation
- Designation_1
- ExpDurationInMonths_1
- Designation_2
- ExpDurationInMonths_2
- Education_1
- EduDurationInMonths_1
- Education_2
- EduDurationInMonths_2
- Total_Exp
- Total_Education
- URL(of Profile)

3

Data Preprocessing

Preprocessing

01 DATA CLEANING

- Data extraction from raw data using string manipulation (regex etc.)

02 TOKENIZATION

- Text data was broken down into list of words called ‘tokens’.
- STOPWORDS were removed

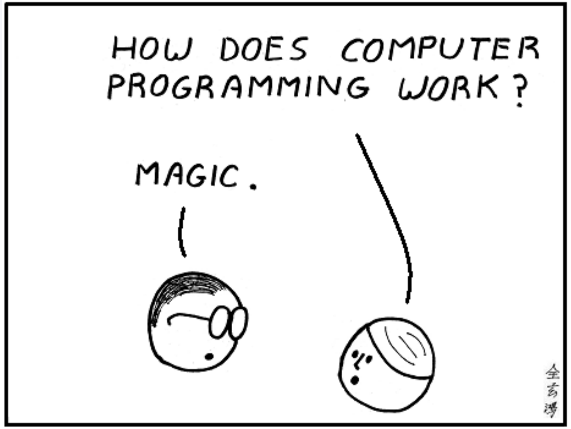
03 LEMMATIZATION

- Root words were found out for all tokens in the text.
- NLTK library is used for lemmatization.

www.bettercartoon.com
©Dimmoe
"Yes! I found it...Now I have to remember what I need it for..."

Now what?

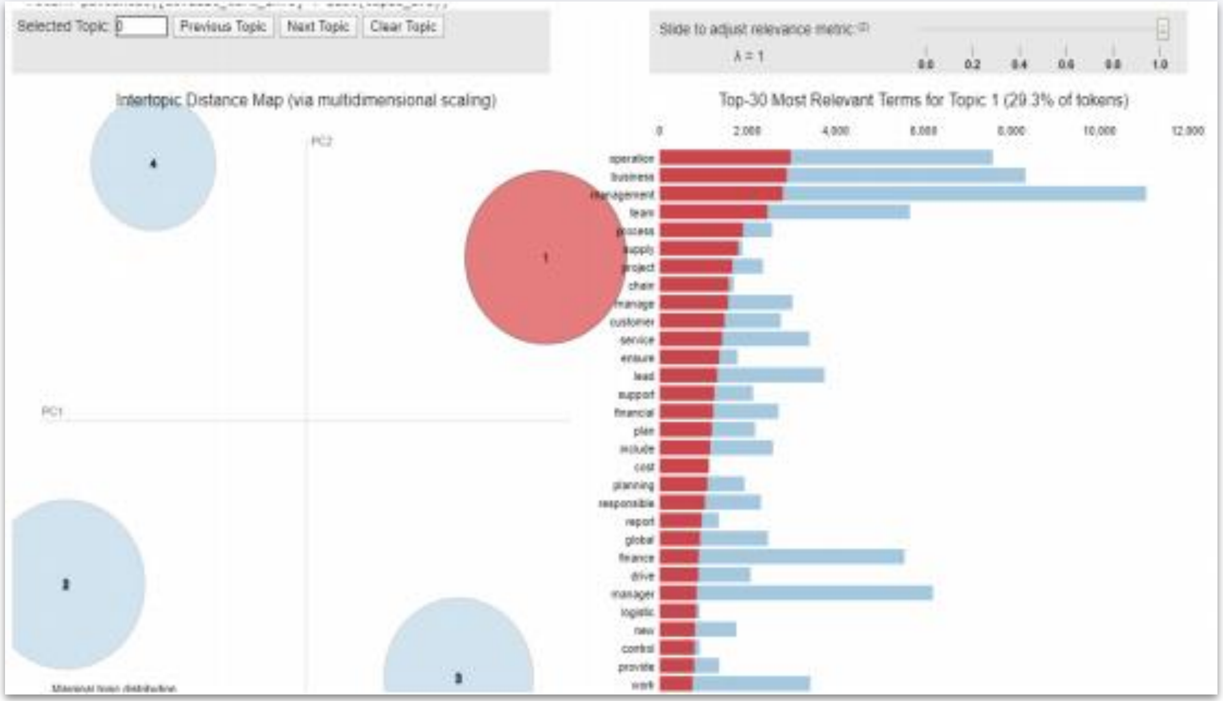
4 Data Modeling



Topic Modelling

- Used sklearn's Latent Dirichlet Allocation(LDA)
- Document-Word Matrix was created using CountVectorizer of sklearn
- Grid Search Optimization was done to find the best model which was used for profiling

Topic Model Output



-HR
-Marketing
-Operations
-Finance

Result : https://docs.google.com/spreadsheets/d/1-hU40jTwhJu9-Q8BrSmlyihM7aPQ2iAgwAg82awqp_E/edit?usp=sharing

Finding the right profile

- Metric used: Jensen-Shannon distance which is the square root of Jensen-Shannon Divergence(JSD).

- JS Distance=(JS Divergence)^(½) is a method of measuring the similarity between two probability distributions (documents in this case).

- The profiles selected are on the basis of the similarity scores between the query and the selected profile.

5

Deployment

- For Integration we used *Flask*
- Requirements?
- How we did it?

Our App Navigation

You search, we recommend

6

Product Marketing

Customer since eternity, "Why your product?"

We do the maths, you take the decision!

“

We don’t believe in 1st,2nd and 3rd because YOU are always 1st!

From the thousands of database, we give you the best search results. Just what you need!

Time:

- Each Page visit in Linkedin starting from searching in the search bar takes 10 sec.

  For 5 profiles you waste 50 sec. (~ 1min.)

We show the same 5 results in a sec.,in one go. You save time!

Money ? Linkedin and other professional networks are costly . We cost less than Linkedin!

Care about organised data? We organise better. Just what you need!

Summary,Name,Experience, and link to know more all in one page. Not 1 but 5 profiles!

We are a new born but we have answers to your "How?"

- Initially want to use *push strategy* for product awareness.

- Segmentation / Target Group:

    - Primary TG: SMEs who can't afford costly recruitment softwares like Linkedin Lite, Naukri premium etc.

    - Secondary TG: Candidates (mostly students) who want to get in touch with people of a particular domain.
        - Remember all those random texts from random people asking for opinions about an institute , career path etc!! Quora here we come!

- Strategy:

Primary Target Group: SMEs and other organisations	Purpose: Revenue earner, partially product awareness.
Secondary Target Group: Students. *<verify students>*	Purpose: Primarily product awareness. Not to be used for revenue.

- Revenue Model: (Initially till product captures a certain market size)

Primary Revenue Earner	SMEs, other organisations
Secondary Revenue Earner	Nil

- Promotion: YouTube ads, Popular web series tie-ups,WOM, our revenue model itself for promotion.
- Competitors: We consider Linkedin our primary competitor- high brand loyalty of users! Besides there are Zoho,Social Recruiting,New radius Search etc.

A needle in a haystack for an MVP and I have read “miles to go before I sleep..”

We found a customer who have shown interest in our product! Yups a ‘Partner’. Eeeaahhh!

P.C 35greeting.com

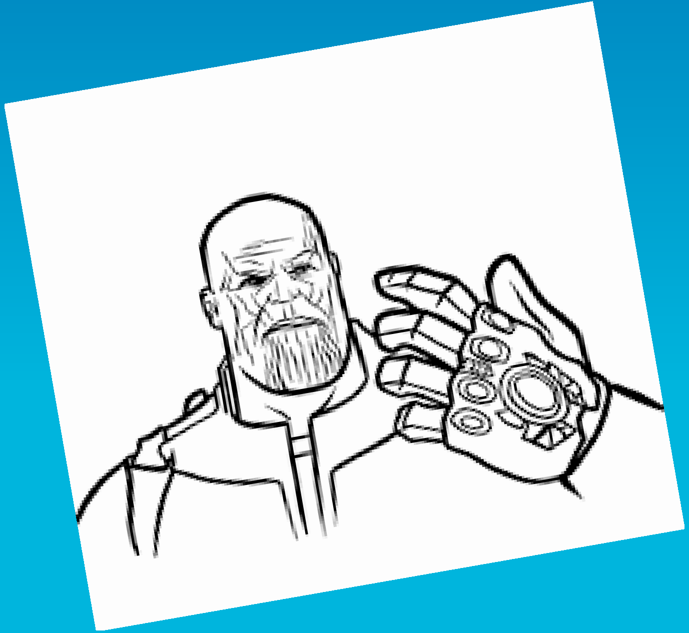
Pricing

Competitors	Plans available(Rs)	Pricing (Rs/month)	Nos. of profiles allowed to visit
Linkedin Recruiter	68840/year	5736.666667	Unlimited
Linkedin Recruiter Lite	9180.97/per month	9180.97	Unlimited
Zoho	1350/month/user	1350	Unlimited
Naukri	83000/3 months	27666.66667	7000

- Planning a penetration pricing initially.Freemium.
- Later on we can make difference in pricing one Basic another Advanced.

7 Conclusion

Step-I : What’s the story so far?



Step-II Limitations?

Step-III : So what next?

We at Forage

SUBHASIS DASGUPTA
MOUMI DAS
SOUMYA BANERJEE
SHILADITYA SWARNAKAR
RAJARSHI CHOWDHURY
SOWMYA KARTHIK
SOUVICK DATTA

Reference Links:

- https://www.glassdoor.com/employers/blog/calculate-cost-per-hire/
- https://www.shrm.org/hr-today/trends-and-forecasting/research-and-surveys/Documents/2016-Human-Capital-Report.pdf
- https://en.wikipedia.org/wiki/Jensen%E2%80%93Shannon_divergence
- https://www.cs.princeton.edu/~runzhey/demo/Topic_Models_I.pdf

Thank you

Any Questions?