# glassdoor Reviews Analysis: Topic Modeling

Sponsored by

SOUTHERN GLAZER'S WINE AND SPIRITS

# Table of Contents

# OBJECTIVES

- The major tasks are to detect words/phrases, and extract these quality texts from any messy data.
- Apply text analysis (Machine Learning) for meaningful patterns from SGWS' s reviews on Glassdoor.
- The relevant benefits are main ideas without having to read large volume of information.

# SCOPE

- The dataset timeline is from September 2016 to March 2019 with 408 reviews in total.
- The final outcome will be a dynamic visualization of popular words and topics.
- Each section of the project will lay out the context for the next step.
- This project consists of three main parts: exploratory analysis, text preprocessing, and Latent Dirichlet Allocation-topic modeling.

# Case Study 1: Pros ⊕

| TOPICS | REVIEWS |
|---|---|
| Topic 1 (19.7%)<br>Benefit, good, pay, perk, nice, training, people → Salary & Benefit | *“Increase base pay and commission rates. Make goals more attainable. “* |
| Topic 2 (19.5%)<br>Freedom, work, schedule, flexible, day, hour → Work Condition | “Start making smart decisions that build your employee loyalties and developing them” |
| Topic 3 (13.1%)<br>Product, hard, supplier, manager, lot, grow → Market Expansion | *“Hire more staff. Get a warehouse in a central location”* |
| Topic 4 (11.4%)<br>Environment, brand, group, network, fun → Work Culture | *“Invest more in the people who do the hard physical labor” “listen to drivers advice”* |

- **Salient Terms:** Great, benefit, good, pay, people, training, opportunity, job, wine, industry, management, flexible, schedule, family, environment, time, free, advancement, growth

# Case Study 2: Cons ⊖

| TOPICS | REVIEWS |
|---|---|
| Topic 1 (90%):<br>Work, management, pay, company, hour, employee, goal, year, day, sale.<br>→ Problems with managers, pay, work hours. | *“Management issues, loads of busy work, high pressure for low pay, you have to hunt down your incentives on your own if you want to get paid.”* |
| Topic 2 & Topic3 (10%) :<br>Much, market, union, stop, account, heavy, communication, sale, delivery.<br>→ smaller group, smaller problems. | *“Lots of last minute "emergencies" and general lack of planning and communication.”* |

- **Sales issue:** time, not, lot, much, sales, manager, team, market, account, customer, communication.
- **Truck driver issue:** union, operation, delivery.

# Case Study 3: Advice to Management

| TOPICS | REVIEWS |
| --- | --- |
| Topic 1 (19.6%)<br>Everyone, commission, think, hard, work, sale.<br>→ Action may be training & treating? | *"Increase base pay and commission rates. Make goals more attainable. "* |
| Topic 2 (13.6%)<br>Develop, open, level, sale, team, stop, company.<br>→ Communicating and developing? | *"Start making smart decisions that build your employee loyalties and developing them"* |
| Topic 3 (12%)<br>Do, not, hire → Hiring and make right decision? | *"Hire more staff. Get a warehouse in a central location"* |
| Topic 4 (11%)<br>Physical, labor, position, hard, driver | *"Invest more in the people who do the hard physical labor" "listen to drivers advice"* |

- **Sentiments**: Not, do, good, physical, hard.
- **Objects**: employee, management, work, business, sale, people, woman.
- **Action**: treat, keep, train, develop, communicate, hire.

# Topic Modeling can be a consistent application to HR analytics including employee retention.

Topic Modeling – What/Why did it happen?

- Process comments, surveys, performance reviews, etc.
- Diagnose/describe the problems.
- Feed text patterns into prediction.

Predictive Modeling- What will happen?

- Classification model is an example.
- Assign a predefined label to each interested group.
- Recommend the right targets for adjustment.

Prescriptive Action – How to make it happen?

- Decision tree is a typical method.
- Optimize a set of decisions to gain the best expected value.
- Provide feedbacks to the diagnosing step.



# EXPLORATORY ANALYSIS

Featured Ratings from Reviewers:

- Bars and trends

Demographic Characteristics:

- Job types, levels, regions

Comments:

- Deleting empty rows/ invalid values

3.3 Rating Trends

57% Recommend to a Friend

81% Approve of CEO

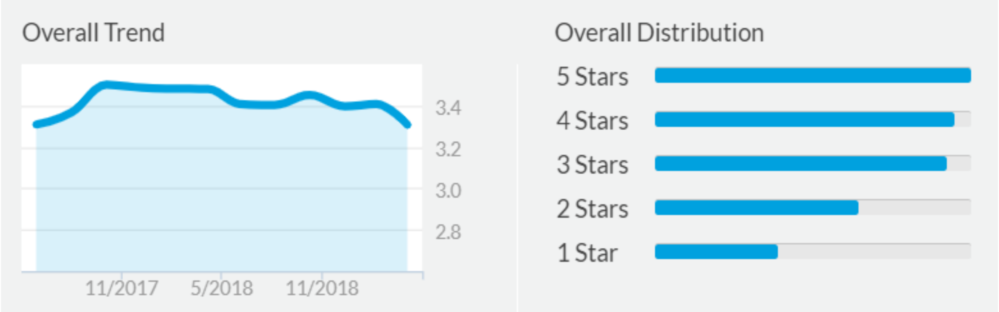
Wayne Chaplin
174 Ratings

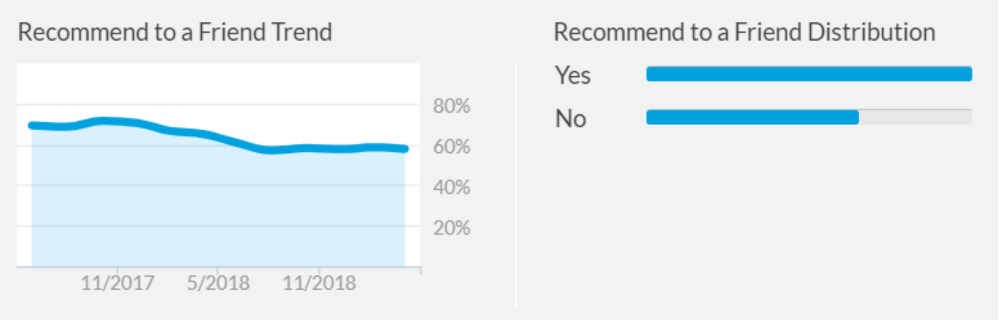
62%
139 Sales Position

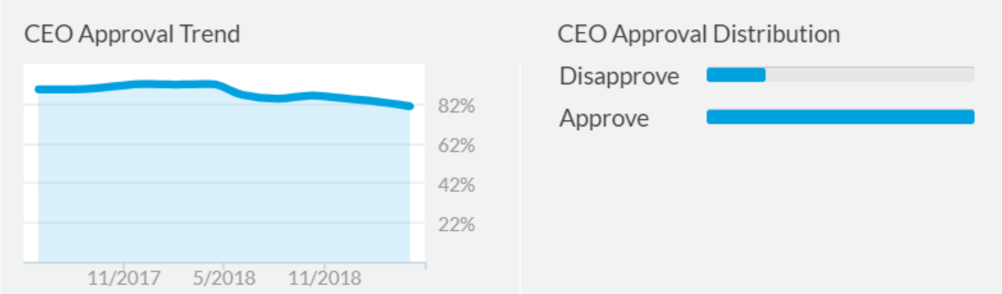
28%
63 Managers

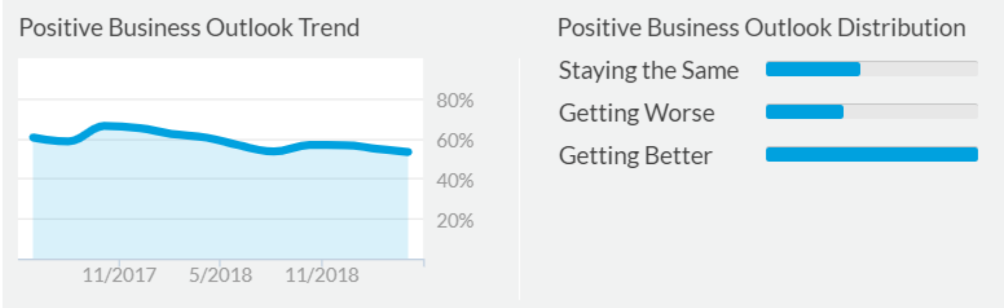
69%
330 Full-time Employees

glassdoor™



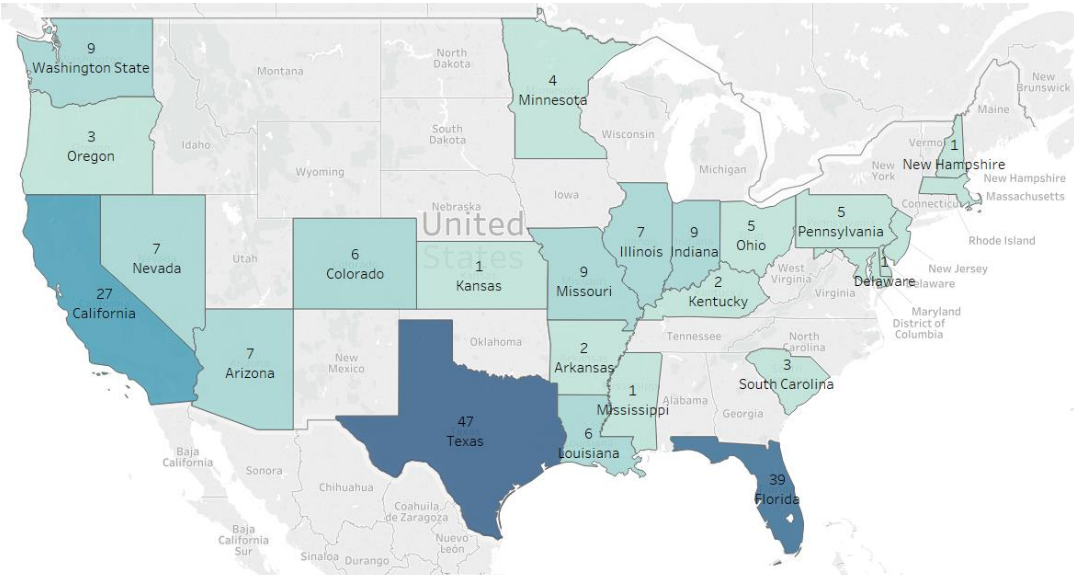



# Background and Focus

- Featured Ratings:
  - Business Outlook, Recommend to friend, CEO Approval, Overall Rating.

- Demographic Characteristics:
  - States: Texas (22%), Florida(18%), California(12%),etc.
  - Job: Sales (62%), Business Analyst (4%), Driver(1%), etc.
  - Level: Managers (28 %).

➢ Reviews are the focus of this project :
  - Pros: 397 (valid texts).
  - Cons: 401.
  - Advice to senior management: 218.




**Pros** "Great Benefits job is extremely physical" (in 36 reviews)

"Great company that allows an amazing work home balance" (in 26 reviews)

**Cons** "Can be a tough work/life balance for some" (in 14 reviews)

"Long hours especially around the holidays" (in 15 reviews)

# TEXT PREPROCESSING

Word Cleaning

- Remove nonsense words, tokenize, and lemmatize.

Phrase Detection

- Pros & Cons.

Text Transformation

- Dictionary & Corpus.

# Word Cleaning – Pre Processing

`'Flexibility\n\n\nHelping Businesses \n\n\nHealth Benefits are good'`

Raw text example

`'Flexibility Helping Businesses Health Benefits are good',`

Remove new line characters

`['flexibility', 'helping', 'businesses', 'health', 'benefits', 'are', 'good']`

Remove punctuations &
Tokenize the words.

`['flexibility', 'help', 'business', 'health', 'benefit', 'good']`

Remove stop word ‘are’ .

`'always busy, good benefits, industry continues to grow ensuring good job security'`

`'continue'` `'ensure'`

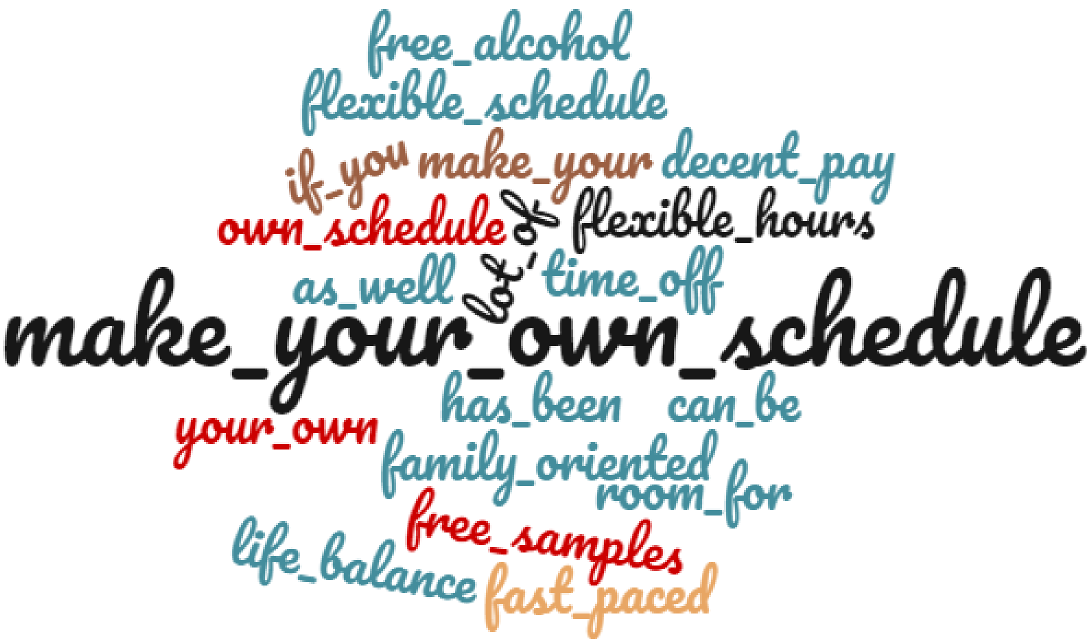
`['always', 'busy', 'good', 'benefit', 'industry', 'continue', 'grow', 'ensure', 'good', 'job', 'security']`

Lemmatization:

- Transforms word to its root.
- Keep noun, adjective, verb, adverb.

# Phrases Detection

PROS




CONS

# TOPIC MODELING:
# Latent Dirichlet Allocation Algorithm

Performance metric

- Number of topics & model quality

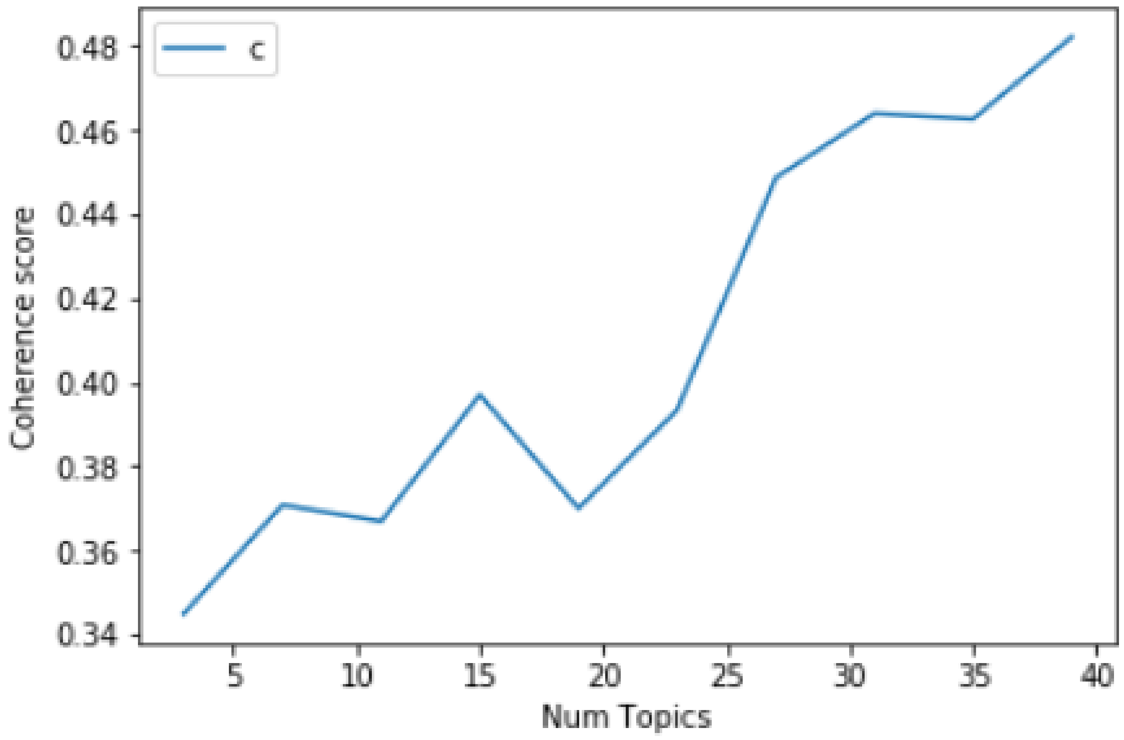
Dynamic visualization

- Glossary
- Findings

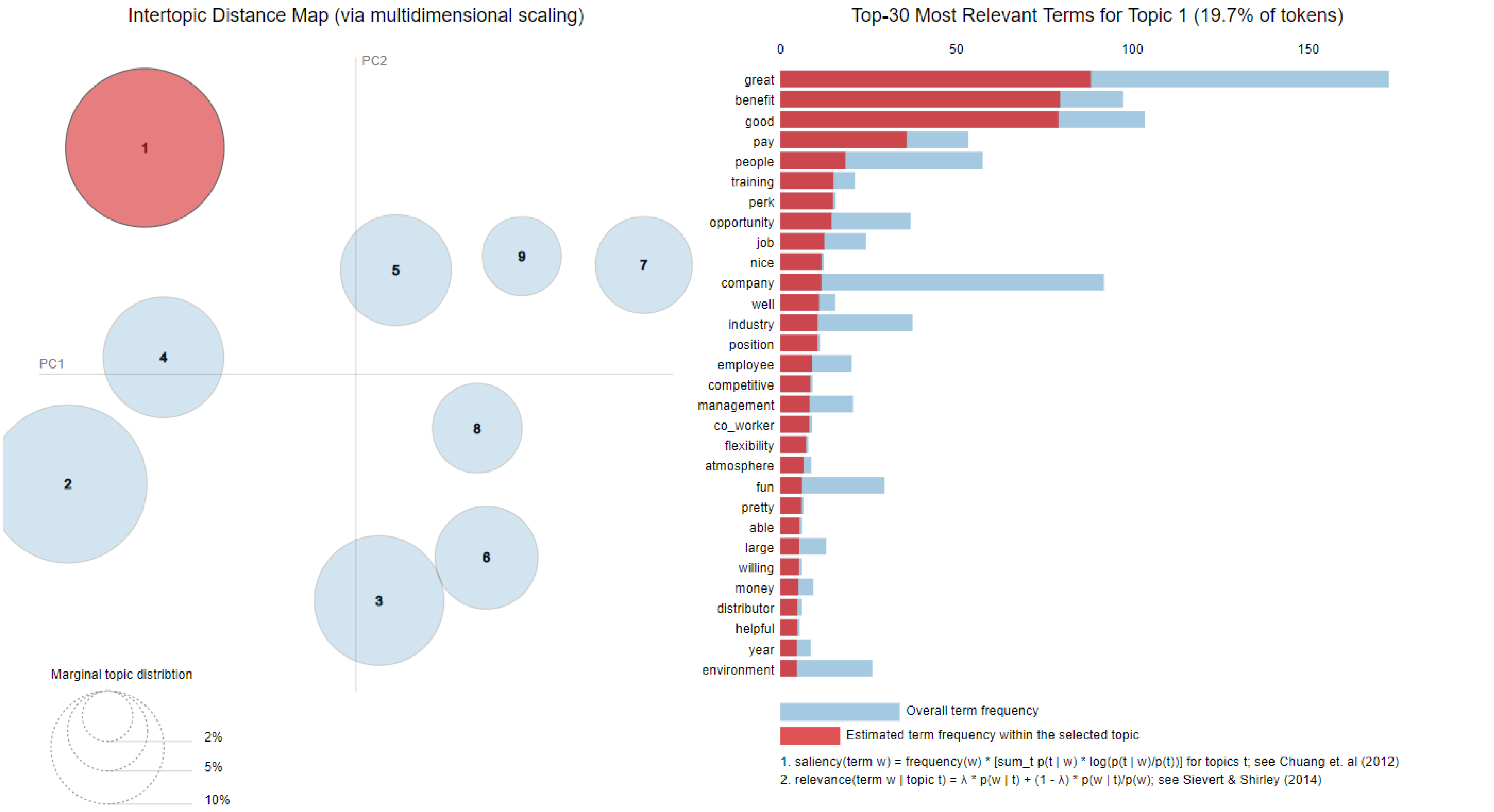
Interpretation

- Recommendations
- Usage Demo: Pros

# Performance Metric




- High coherence score means the words in each topics are more relevant to each other.
- However, the trade-off of too many topics is difficulty in interpretation and visualization.
- A rule of thumb is to test and try with number of topics manually until all the topic bubbles are evenly dispersed or not overlapped.
- In this example with Pros, the decent number of topics is 9. Human intelligence still counts.


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (19.7% of tokens)
0
50
100
150
great
benefit
good
pay
people
training
perk
opportunity
job
nice
company
well
industry
position
employee
competitive
management
co_worker
flexibility
atmosphere
fun
pretty
able
large
willing
money
distributor
helpful
year
environment
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

## Dynamic Visuals: Glossary

- Intertopic Distance Map
- Marginal Topic Distribution
- Overall Term Frequency
- Top Salient Terms
- Lambda Metric (Λ)

## Findings

- Topics are unlabeled, so we need to find the words for labelling.
- Do not care much about overlapping or tiny topics. They are noise.
- This LDA algorithm will make more sense in more specified topics.

## Usage Recommendation

- Approaching salient terms to set up an overall context for further investigation of topics.
- Drafting categories based on work themes or part-of-speech (noun, verbs, adjective).
- Investigating topics that dominate at least 50%. Closer bubbles may share some meanings.
- Adjusting relevance metric $\lambda$ to balance popular terms with unique ones, and pick the least changed. A suggested metric is between 0.4-0.6
- Refer some chosen words to the original reviews to better understand some syntax.
- Keeping a context in mind before diving further. What type of reviews? Who wrote them?

# Appendix

# How many topics are enough?

```
# Build LDA model
lda_model = gensim.models.ldamodel.LdaModel(corpus=corpus,id2word=id2word,
                                           random_state=100, update_every=1,
                                           chunksize=100,passes=10, num_topics=9,
                                           alpha='auto',per_word_topics=True )

# Compute Coherence Score
coherence_model_lda = CoherenceModel(model=lda_model, texts=pros_lemmatized,
                                    dictionary=id2word, coherence='c_v')
coherence_lda = coherence_model_lda.get_coherence()

# Visualization
limit=40; start=3; step=4;
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

QUESTION ?