BADS7204 TEXT ANALYTICS AND NATURAL LANGUAGE PROCESSING

Job Description Skill Extraction

NLP Topic Modeling with Latent Dirichlet Allocation (LDA) using Python GENSIM and NLTK

AGENDA

01

OBJECTIVE

02

DATA SET

03

PROCESS

04

EVALUATE

OBJECTIVE

JOB DESCRIPTION SKILLS EXTRACTION TO SUPPORT RESUME MANAGEMENT BASE ON THE REQUIREMENTS OF AN ONGOING RECRUITMENT MANAGEMENT SYSTEM

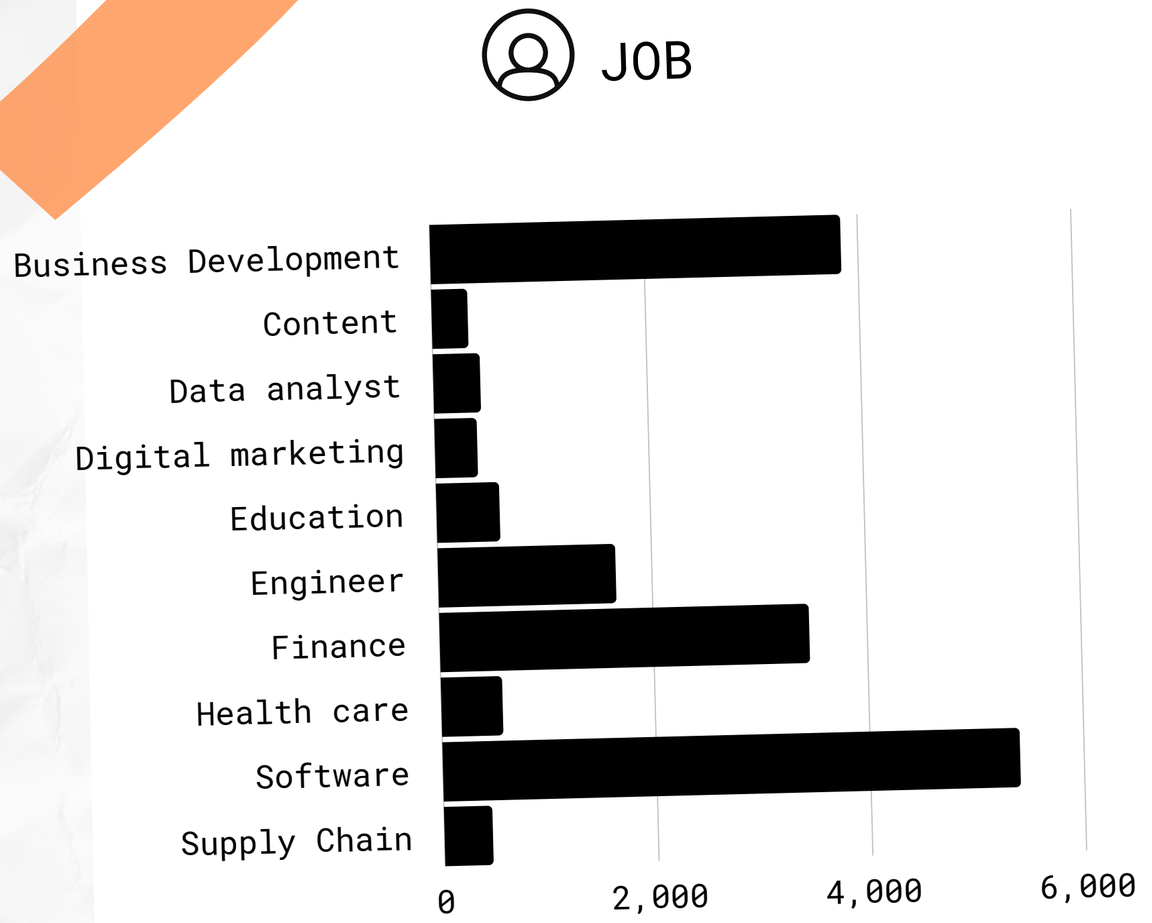
---

our model helps screening job description with in no time it makes process easy, reduce time of job application and increase the chance of getting job.

DATA SET

LinkedIn

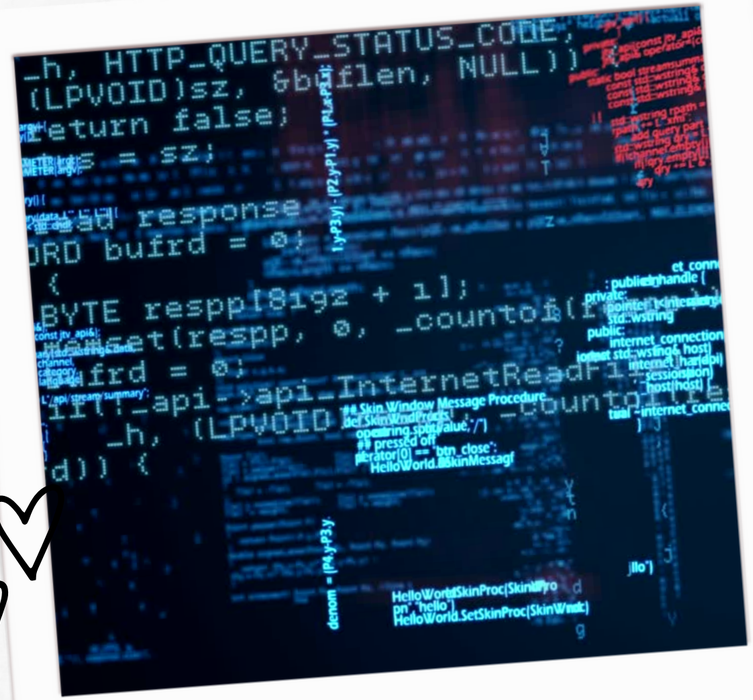
	group	Title	FullDescription
0	Business development	Business Development and Sales Executive	Our client has an exciting opportunity for a n...
1	Business development	Business Development BPO Supply Chain ERP/SAP	New Business Sales Specialist to work for our ...
2	Business development	Business Development Manager (Supply Chain)	BUSINESS DEVELOPMENT MANAGER SUPPLY CHAIN Hert...
3	Business development	Business Development	Business Development Full Time Temp to Perm ...
4	Business development	Business Development Director New Media Up t...	An up and coming Search agency is in need of a...



DATA SET

Selected Top 10 Most Interesting Job fields in 2021

OUR PROCESS



GREAT COMPANY

COLLECT DATA

LinkedIn

DATA PREPARATION

We are here!

MODEL

LDA

EVALUATE

Perplexity
Coherence Score

Data Preparation

- [x] Removing Stop Words
- [x] Remove Punctuation
- [x] Word tokenize and lowercase
- [x] make bigrams & trigrams
- [x] TEXT LEMMENTATION

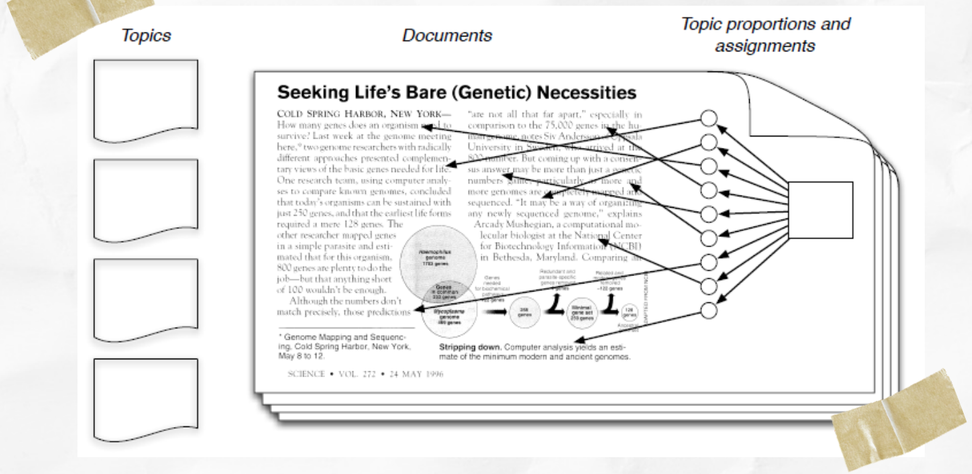
CODE

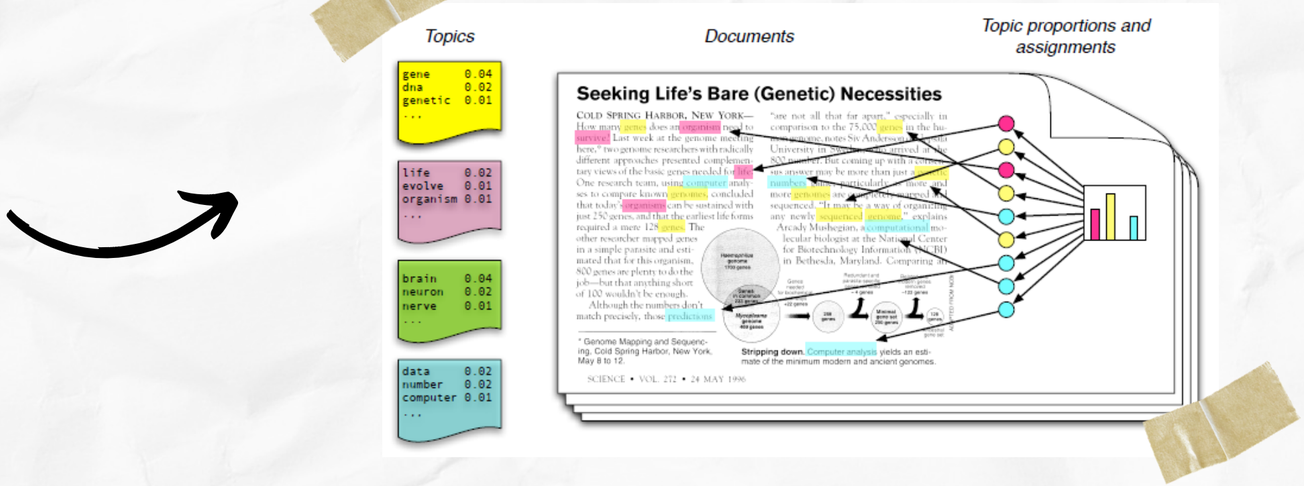
```
#FUNCTION HERE

STOP_WORDS = STOPWORDS.WORDS('ENGLISH')

STOP_WORDS.EXTEND(['WORK', 'ROLE', 'PROJECT', 'SKILL',
'CLIENT' ,'EXPERIENCE'
,'OPPORTUNITY','CANDIDATE','JOB','APPLICATION','INCLUDE'
,'PAGE','RECIEVE','EMBED','NEW','LOOK','WWW'])

DEF SENT_TO_WORDS(SENTENCES):
 FOR SENTENCE IN SENTENCES:
 # DEACC=TRUE REMOVES PUNCTUATIONS
 YIELD(GENSIM.UTILS.SIMPLE_PREPROCESS(STR(SENTENCE),
DEACC=TRUE))

DEF REMOVE_STOPWORDS(TEXTS):
 RETURN [[WORD FOR WORD IN SIMPLE_PREPROCESS(STR(DOC))
 IF WORD NOT IN STOP_WORDS] FOR DOC IN TEXTS]
```

```
def make_bigrams(texts):
 return [bigram_mod[doc] for doc in texts]

def make_trigrams(texts):
 return [trigram_mod[bigram_mod[doc]] for doc in
texts]

def lemmatization(texts, allowed_postags=['NOUN',
'ADJ', 'VERB', 'ADV']):
 """https://spacy.io/api/annotation"""
 texts_out = []
 for sent in texts:
 doc = nlp(" ".join(sent))
 texts_out.append([token.lemma_ for token in doc if
token.pos_ in allowed_postags])
 return texts_out
```

M O D E L

Latent Dirichlet Allocation (LDA) is a Latent Probabilistic Model used in Topic Modeling. This special topic is a hidden topic or theme.





CODE

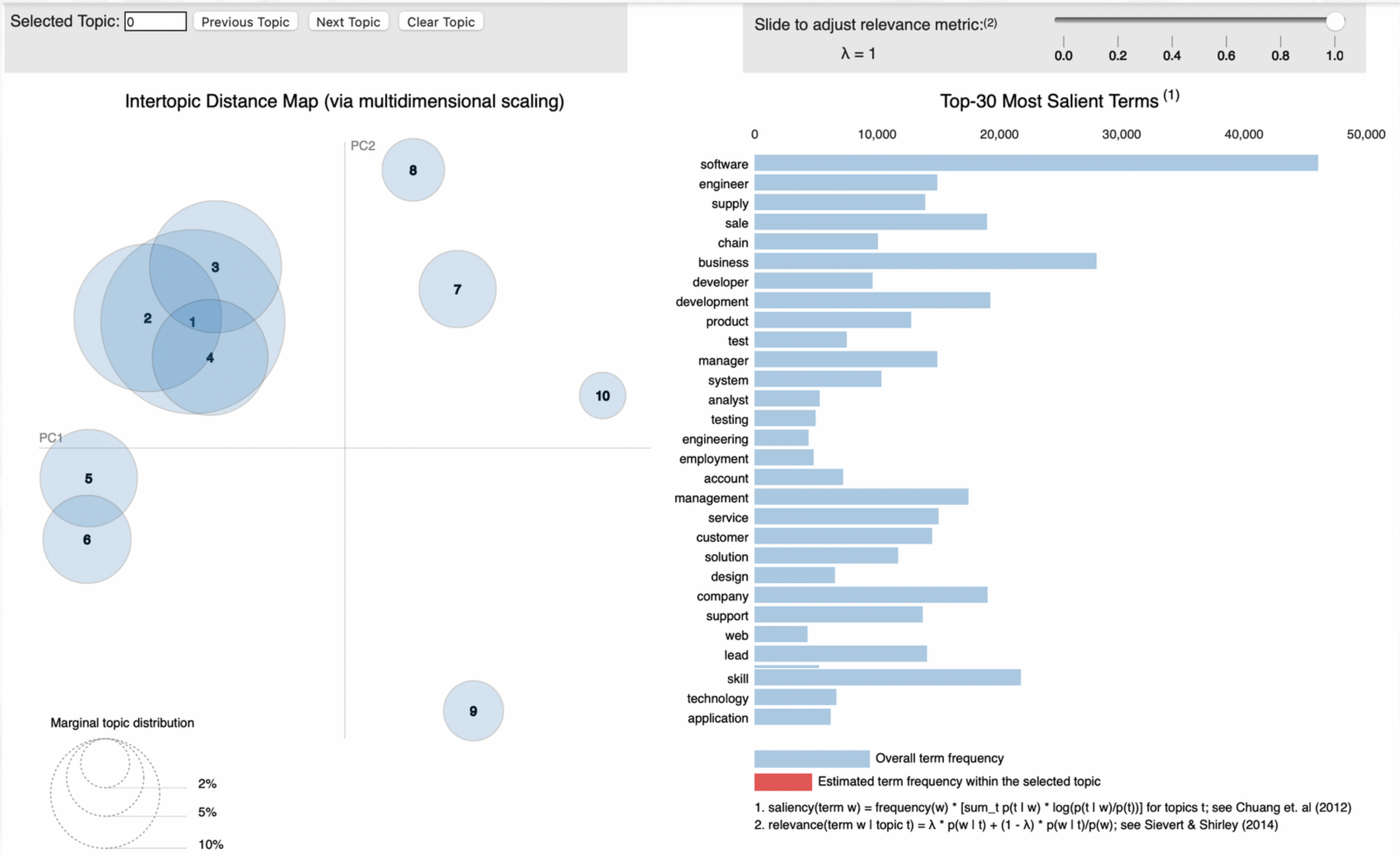
```
# NUMBER OF TOPICS
NUM_TOPICS = 10
# BUILD LDA MODEL
LDA_MODEL_4 =
GENSIM.MODELS.LDAMULTICORE(CORPUS=CORPUS,
                           ID2WORD=ID2WORD,
                           NUM_TOPICS=NUM_TOPICS,
                           RANDOM_STATE=100,
                           EVAL_EVERY=1,
                           CHUNKSIZE=100,
                           PASSES=10,
                           PER_WORD_TOPICS=TRUE)
```

RESULT

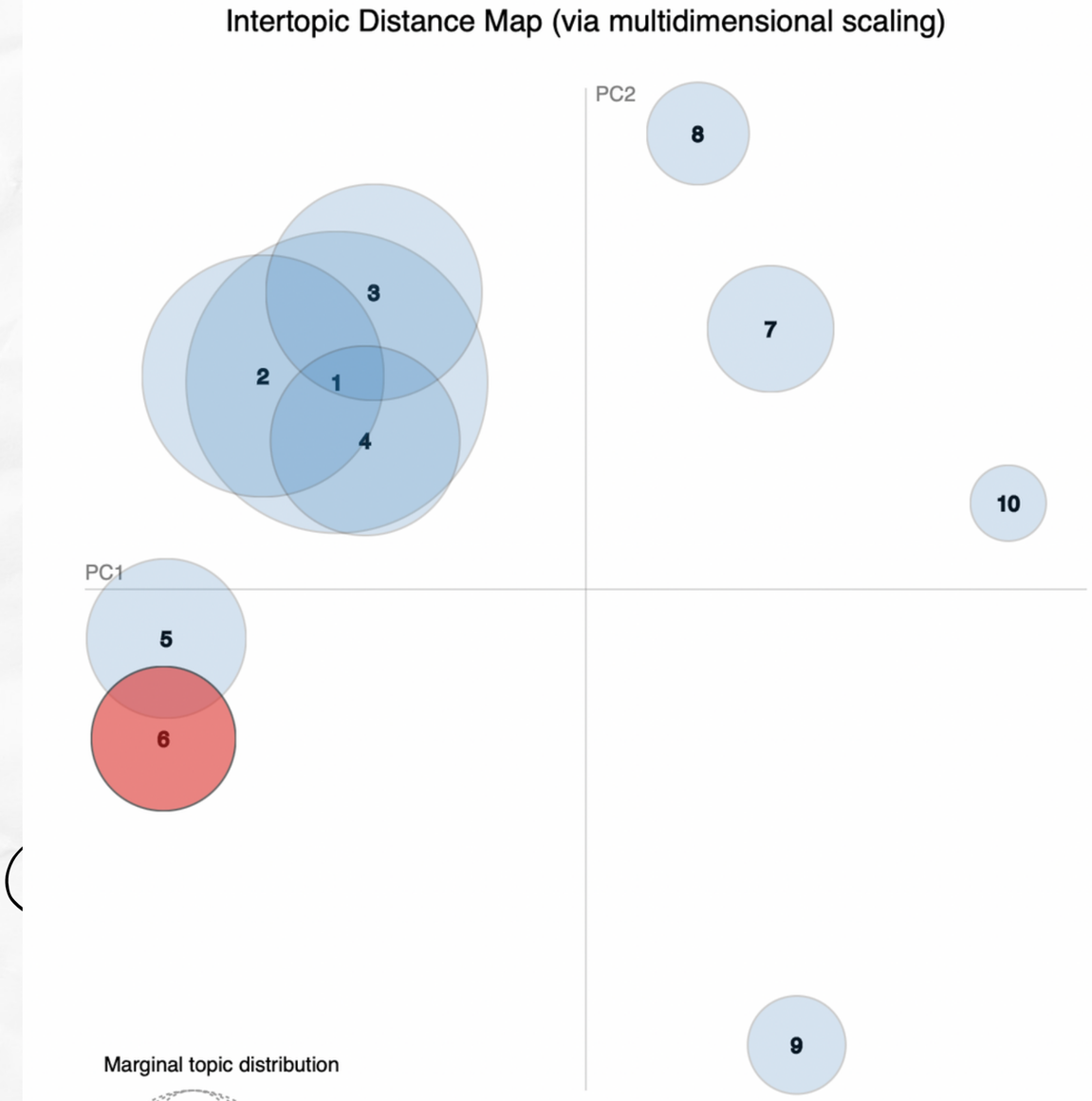
```
[(0,
  '0.075*"sale" + 0.039*"business" + 0.037*"software" + 0.028*"manager" +
'
  '0.028*"account" + 0.026*"solution" + 0.026*"service" +
0.018*"consultant" + '
  '0.016*"oracle" + 0.016*"management"'),
 (1,
  '0.031*"business" + 0.024*"management" + 0.021*"team" + 0.016*"manager"
+ '
  '0.015*"lead" + 0.015*"manage" + 0.011*"strong" + 0.011*"development" +
'
  '0.010*"process" + 0.010*"skill"'),
 (2,
  '0.078*"software" + 0.065*"developer" + 0.037*"development" +
0.029*"web" + '
  '0.026*"technology" + 0.023*"net" + 0.020*"skill" + 0.017*"agile" + '
  '0.016*"product" + 0.015*"develop"'),
 (3,
  '0.019*"company" + 0.014*"team" + 0.014*"look" + 0.013*"work" + '
  '0.012*"excellent" + 0.012*"skill" + 0.012*"base" + 0.011*"successful"
+ '
  '0.010*"join" + 0.010*"career"'),
 (4,
  '0.047*"analyst" + 0.043*"employment" + 0.024*"business" +
0.024*"group" + '
  '0.023*"applicant" + 0.021*"agency" + 0.019*"supply" + 0.019*"sql" + '
  '0.018*"datum" + 0.017*"application"'),
```

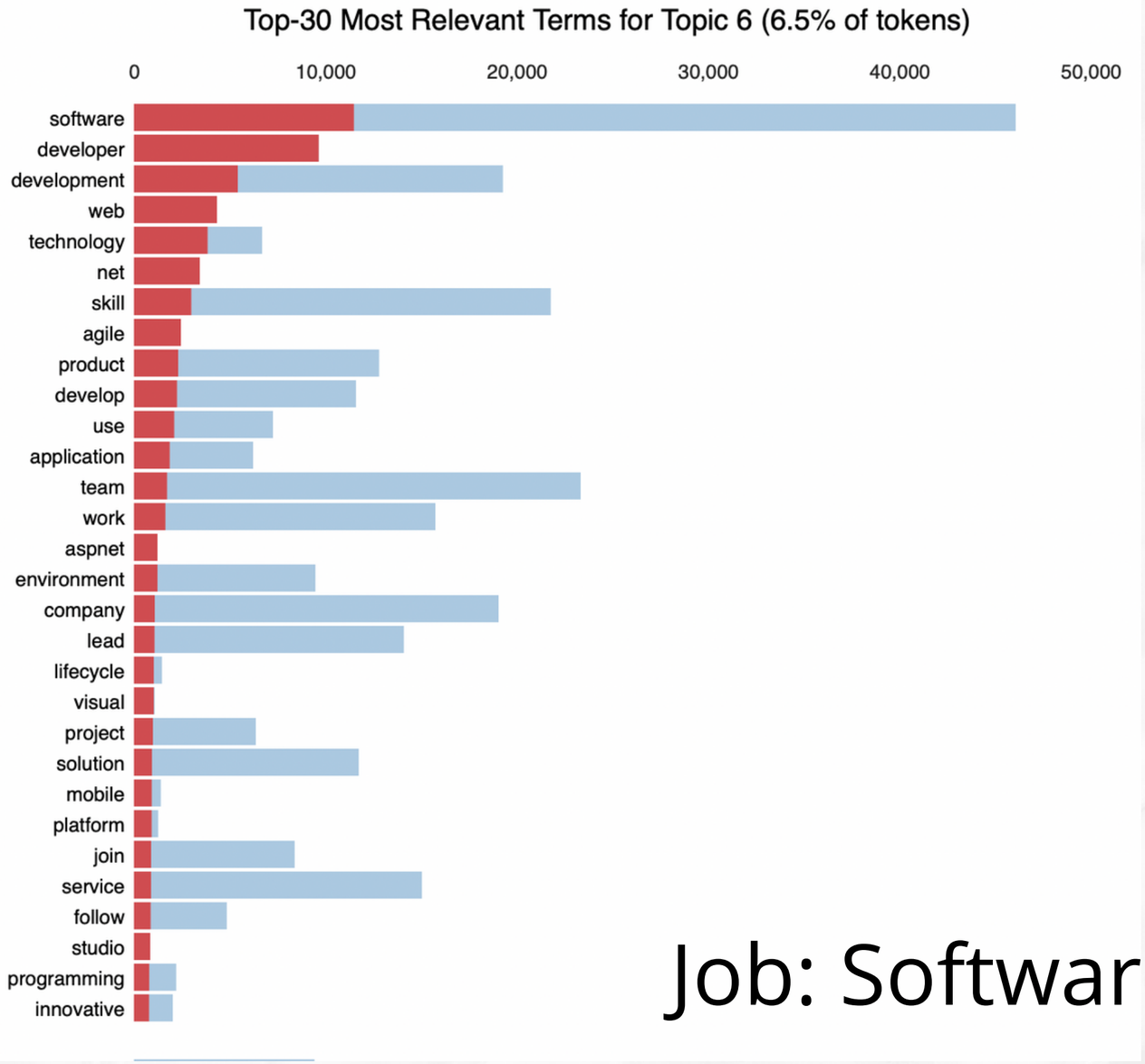
```
(5,
  '0.140*"software" + 0.042*"test" + 0.036*"development" +
0.033*"system" + '
  '0.029*"design" + 0.020*"technical" + 0.019*"embed" +
0.016*"server" + '
  '0.014*"knowledge" + 0.014*"skill"'),
 (6,
  '0.219*"engineer" + 0.065*"engineering" + 0.051*"testing" +
0.049*"product" '
  '+ 0.020*"control" + 0.017*"automation" + 0.017*"industry" + '
  '0.015*"automotive" + 0.015*"equipment" + 0.014*"company"'),
 (7,
  '0.036*"architecture" + 0.028*"training" + 0.022*"deployment" + '
  '0.019*"service" + 0.018*"care" + 0.017*"people" +
0.016*"assurance" + '
  '0.013*"support" + 0.013*"health" + 0.011*"hour"'),
 (8,
  '0.023*"customer" + 0.021*"support" + 0.018*"ensure" +
0.014*"supplier" + '
  '0.013*"skill" + 0.011*"system" + 0.011*"require" +
0.011*"process" + '
  '0.010*"supply" + 0.010*"maintain"'),
 (9,
  '0.190*"chain" + 0.184*"supply" + 0.038*"logistic" +
0.038*"procurement" + '
  '0.017*"manufacture" + 0.016*"application" + 0.015*"manufacturing"
+ '
  '0.015*"lead" + 0.015*"purchase" + 0.014*"day"')]
```

RESULT


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms (1)
0
10,000
20,000
30,000
40,000
50,000
software
engineer
supply
sale
chain
business
developer
development
product
test
manager
system
analyst
testing
engineering
employment
account
management
service
customer
solution
design
company
support
web
lead
skill
technology
application
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)

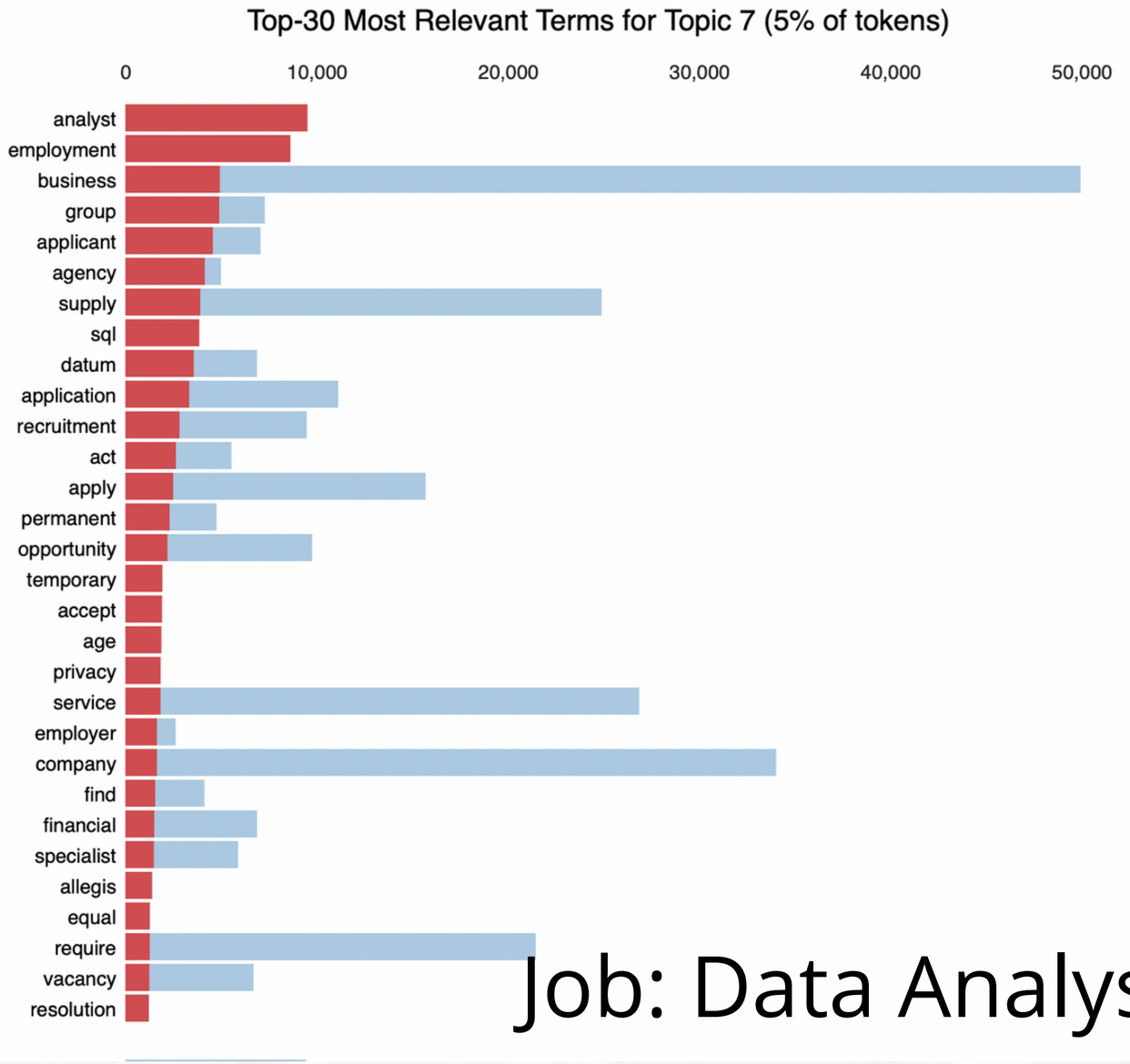
RESULT





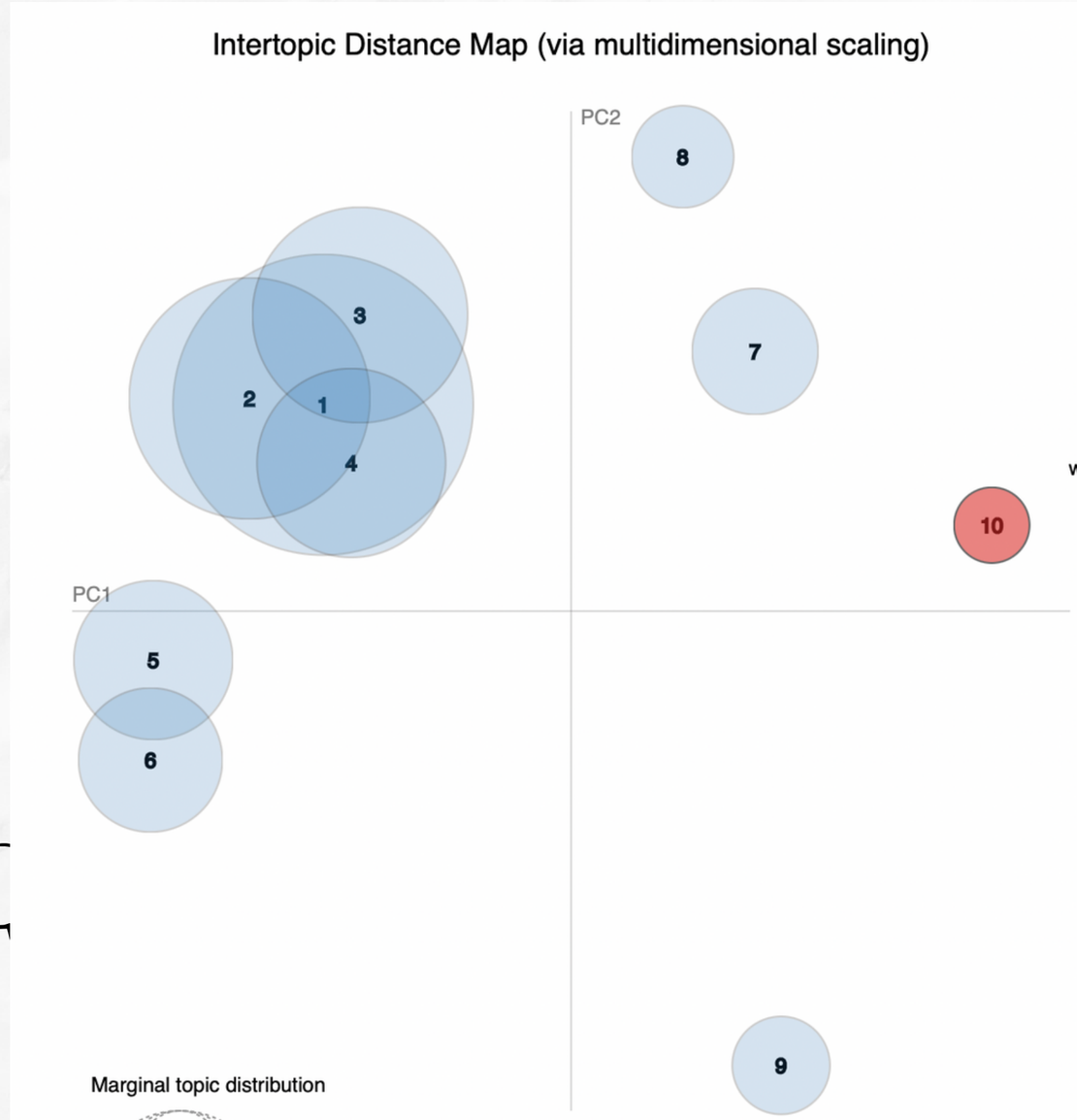
Job: Software

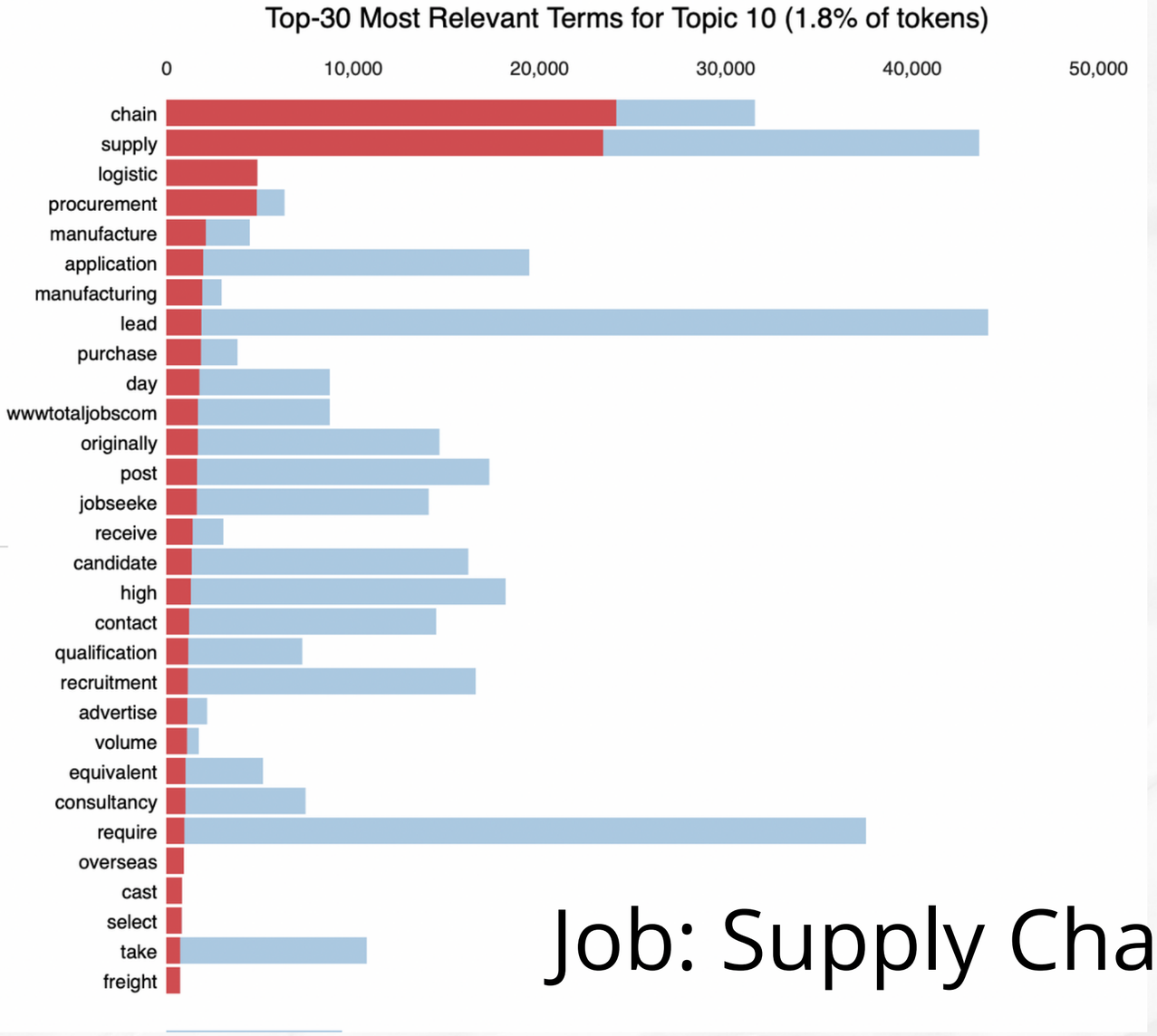
RESULT



Job: Data Analyst

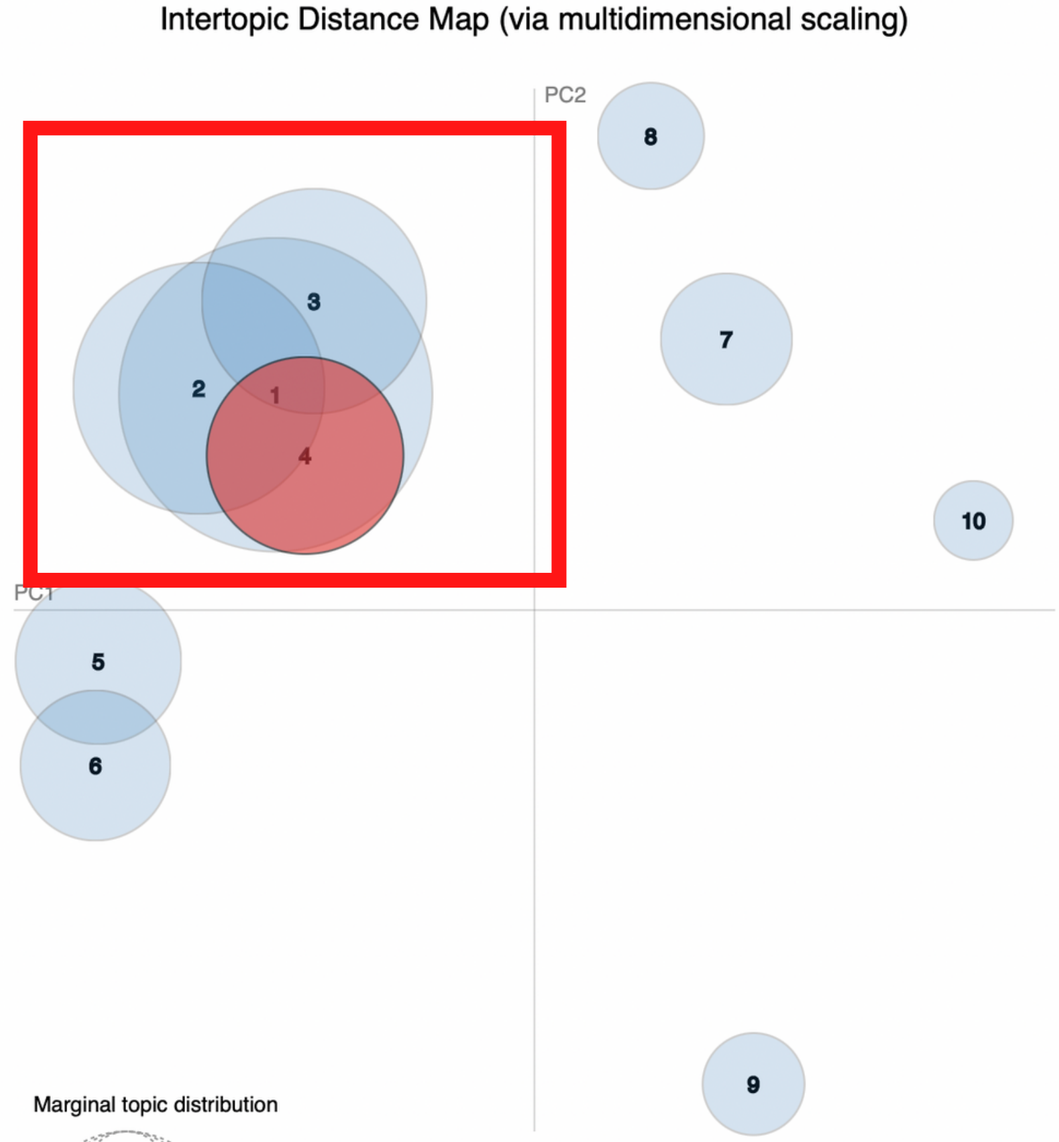
RESULT

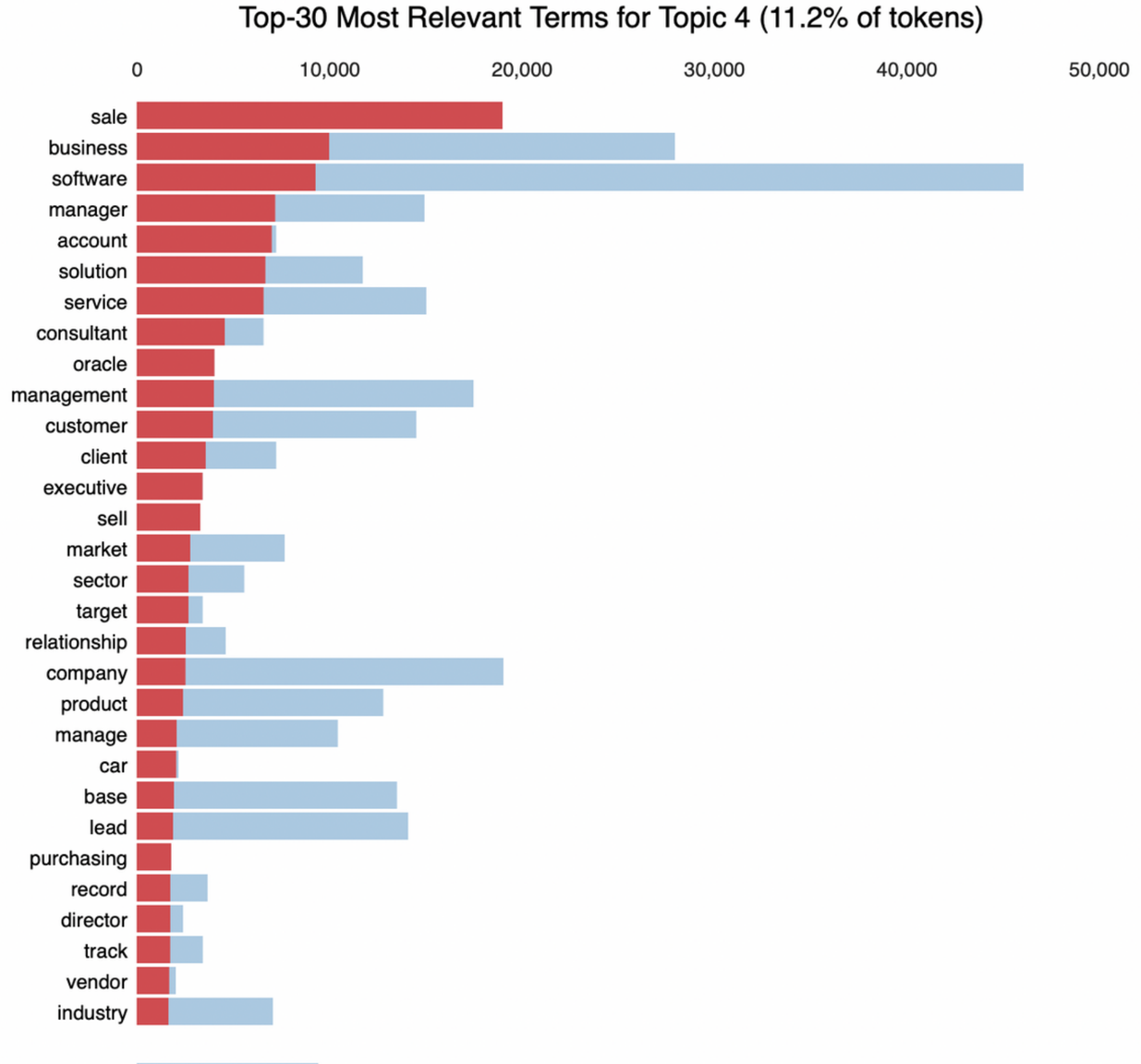




Job: Supply Chain

RESULT


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
Marginal topic distribution

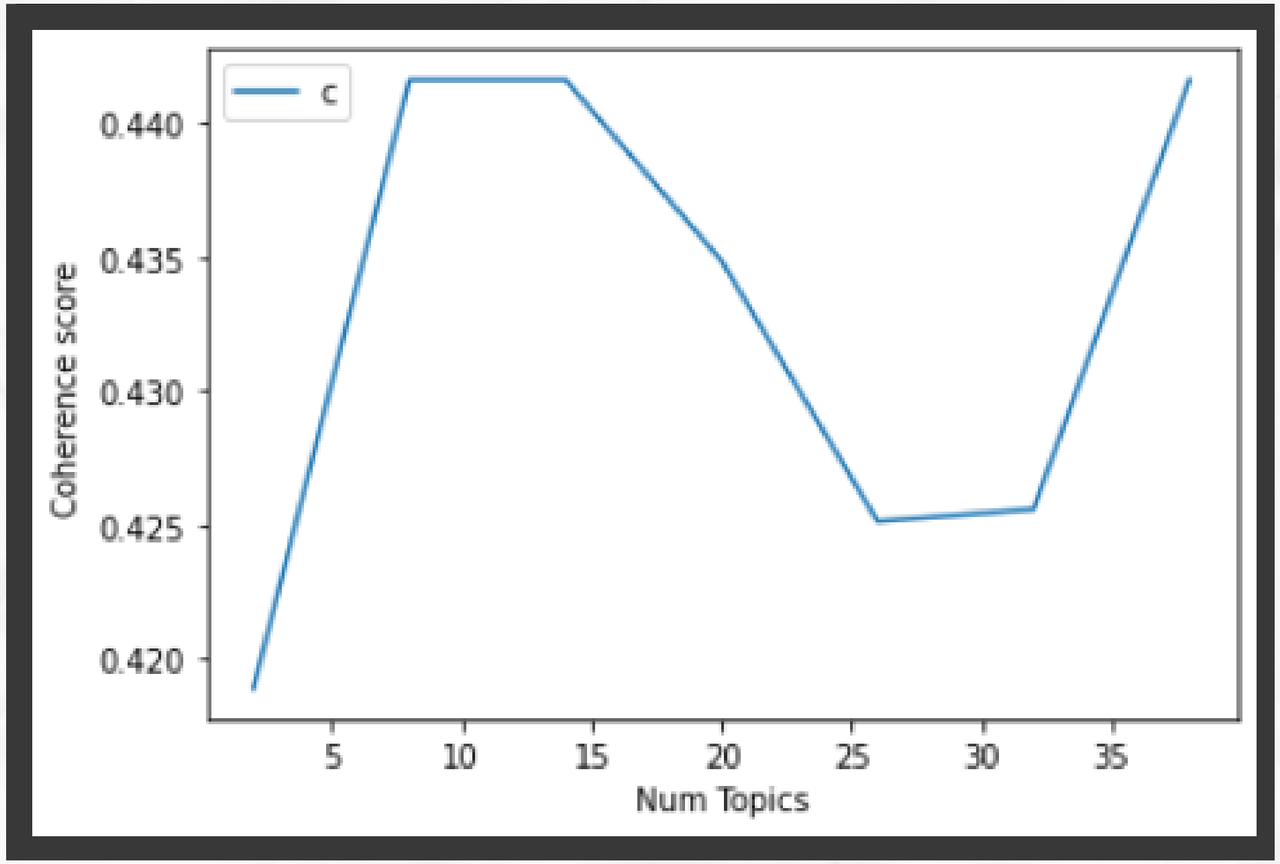

Top-30 Most Relevant Terms for Topic 4 (11.2% of tokens)
0
10,000
20,000
30,000
40,000
50,000
sale
business
software
manager
account
solution
service
consultant
oracle
management
customer
client
executive
sell
market
sector
target
relationship
company
product
manage
car
base
lead
purchasing
record
director
track
vendor
industry

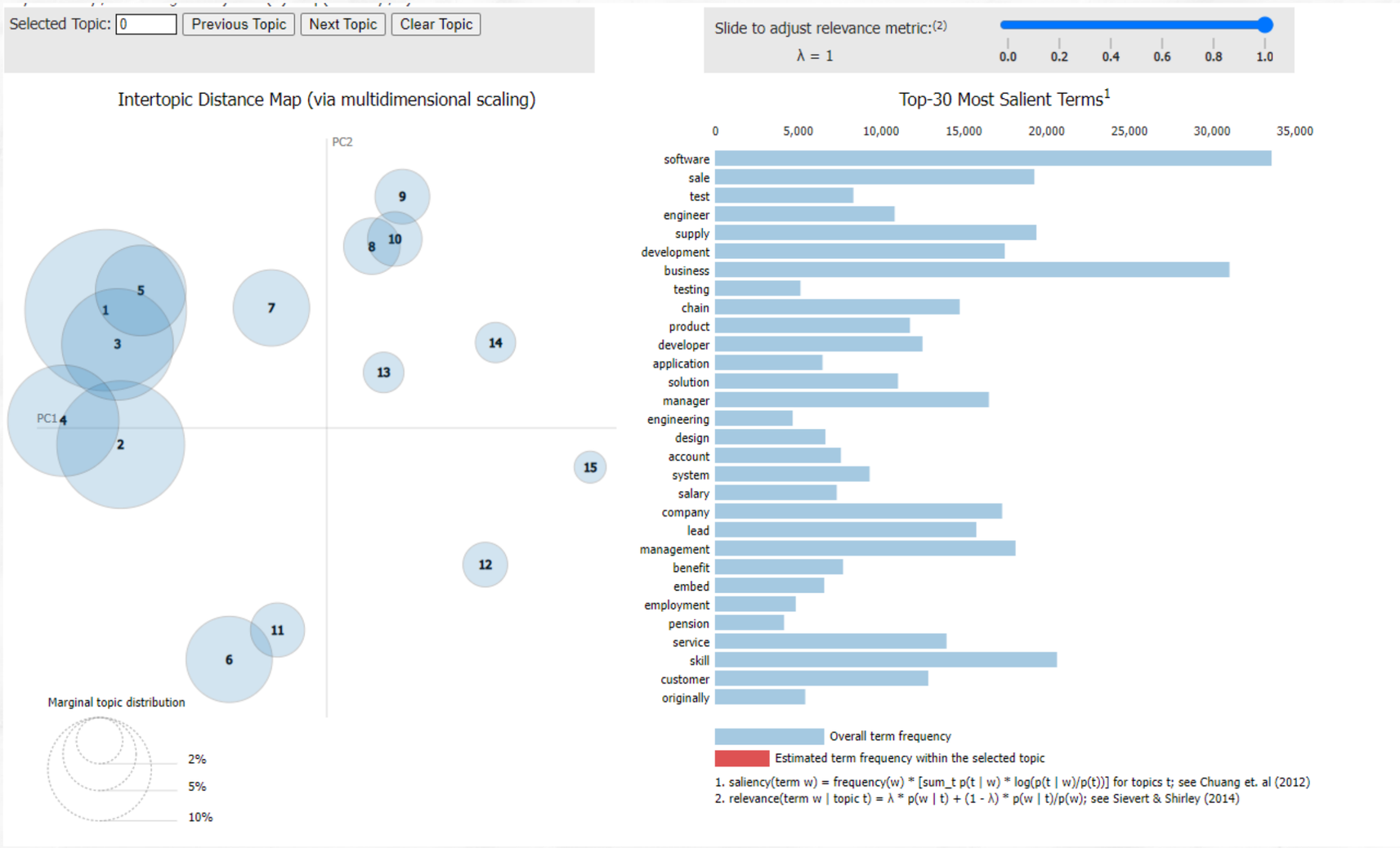
EVALUATE

Perplexity: -7.191671374435629

Coherence Score: 0.445205452797316 5

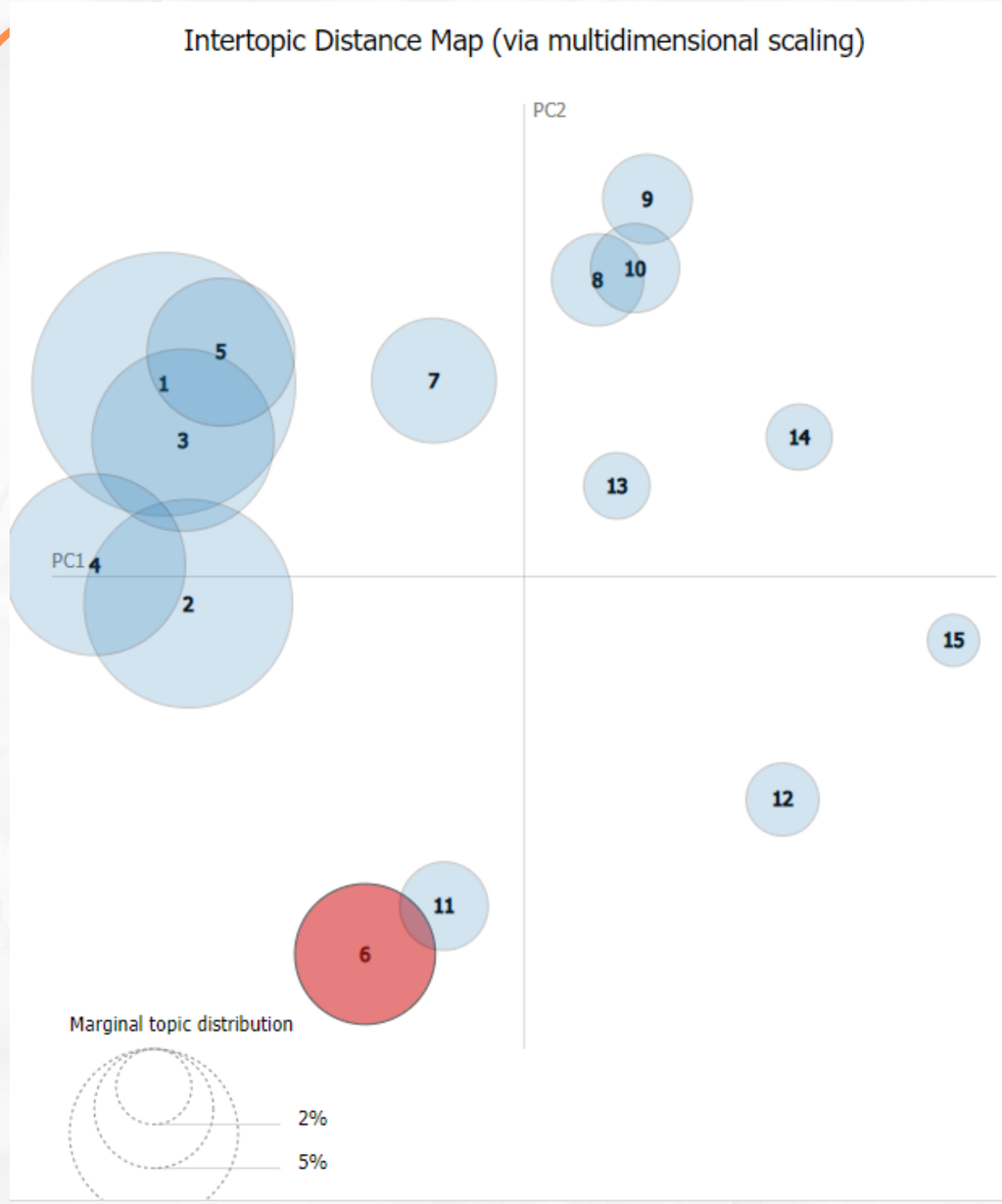
```
# SHOW GRAPH
LIMIT=40; START=2; STEP=6;
X = RANGE(START, LIMIT, STEP)
PLT.PLOT(X, COHERENCE_VALUES)
PLT.XLABEL("NUM TOPICS")
PLT.YLABEL("COHERENCE SCORE")
PLT.LEGEND(("COHERENCE_VALUES"), LOC='BEST')
PLT.SHOW()
```

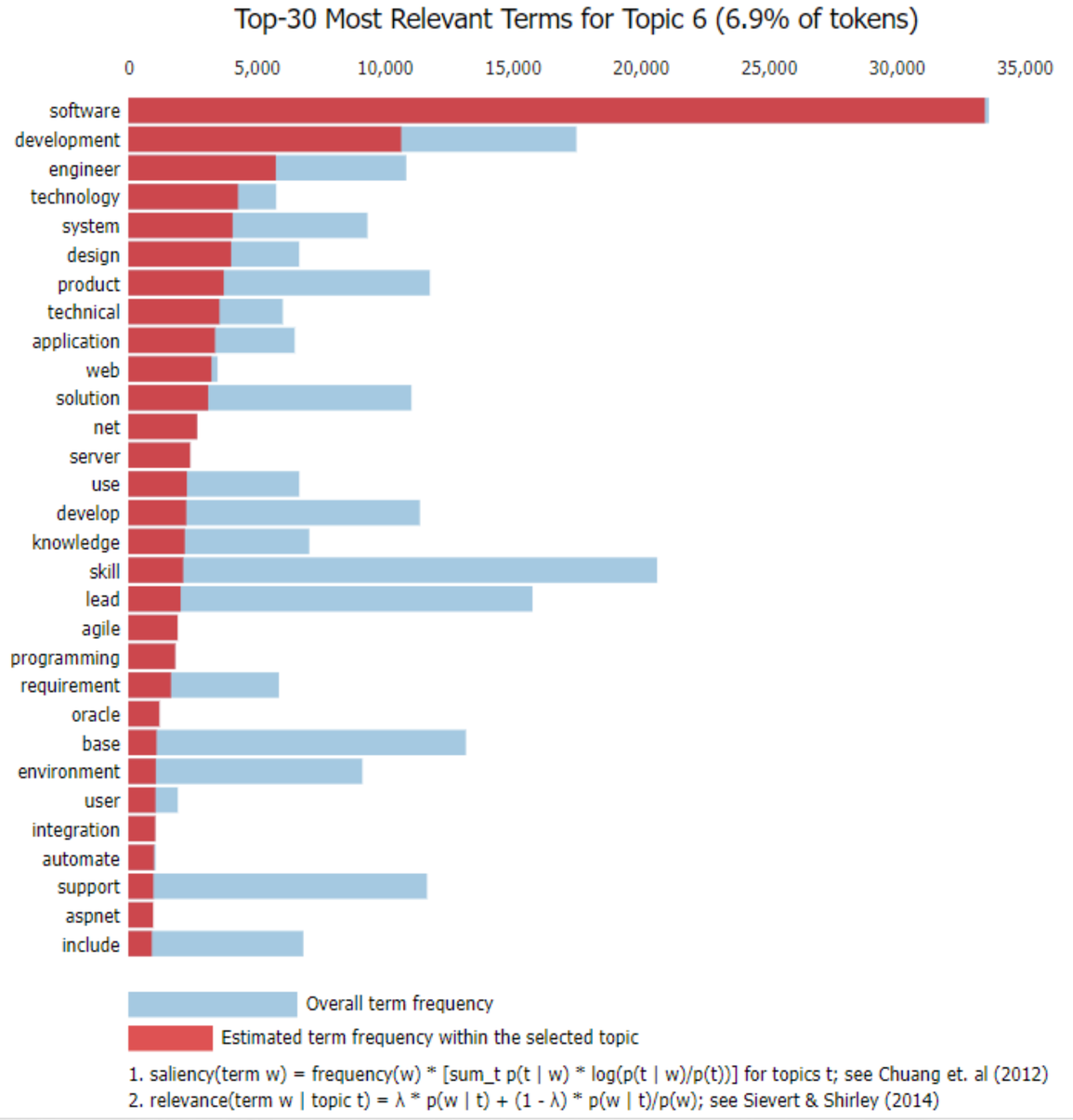




Perplexity: -7.254067944100371

Coherence Score: 0.4452054527973165


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Marginal topic distribution
2%
5%


Top-30 Most Relevant Terms for Topic 6 (6.9% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
software
development
engineer
technology
system
design
product
technical
application
web
solution
net
server
use
develop
knowledge
skill
lead
agile
programming
requirement
oracle
base
environment
user
integration
automate
support
aspnet
include
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)

The Team

Voramate Plodprong
6310422028

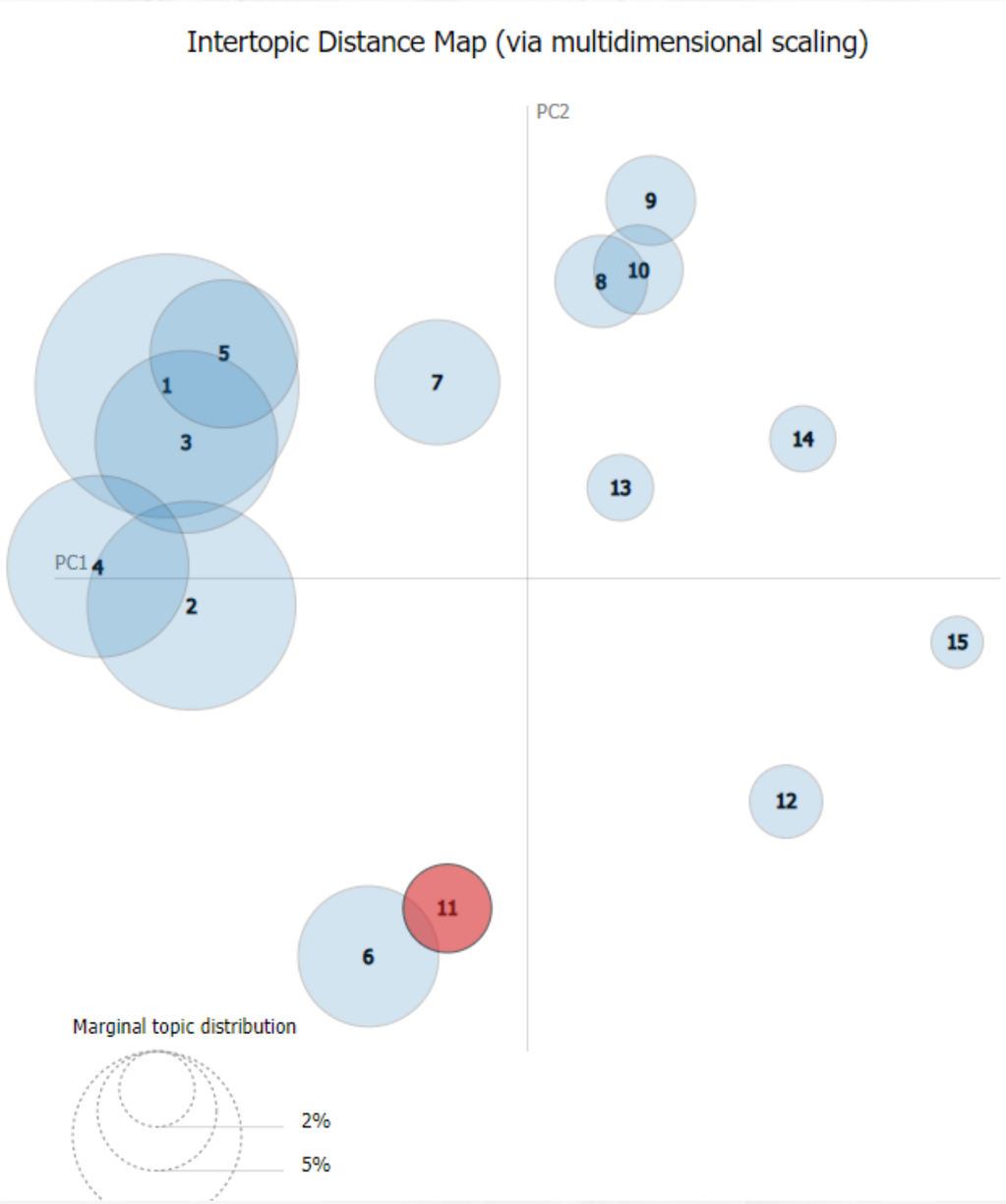
Suphanat Thaiprasit
6310422034

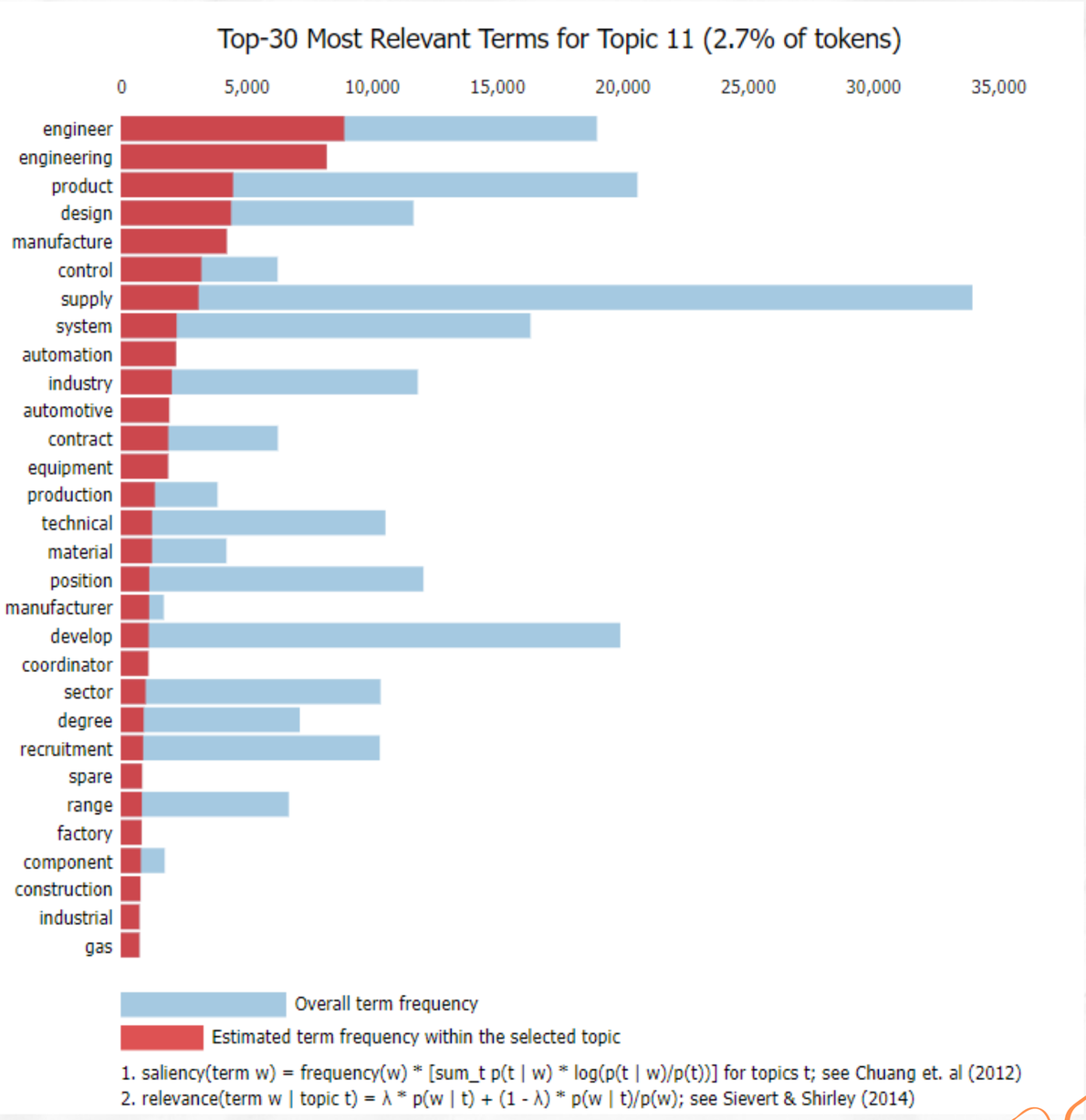
Rakchanok Thongkumpan
6310422039

Varattaya Rojanarachneekorn
6310422044

Piyaboon Kunakornjittirak
6310422047

THANK YOU!


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
9
10
8
5
1
3
7
14
13
4
2
15
12
11
6
Marginal topic distribution
2%
5%


Top-30 Most Relevant Terms for Topic 11 (2.7% of tokens)
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
engineer
engineering
product
design
manufacture
control
supply
system
automation
industry
automotive
contract
equipment
production
technical
material
position
manufacturer
develop
coordinator
sector
degree
recruitment
spare
range
factory
component
construction
industrial
gas
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)