Data Science Bootcamps with reddit

Business Problem

- Imagine yourself as a marketer for a data science boot camp.

  What might you care about?

- How do you know what people are honestly interested in since reviews are public.

Reddit

- Site that hosts any topic imaginable
- High utility since there are numerous subreddits discussing data science, boot camps, and data science boot camps.
- Users are anonymous so opinions shared are more likely to be honest (increase in the validity of comments).
- Good site for asking questions and getting answers. StackOverflow for everything else.
- https://www.reddit.com/r/datascience/search?q=data+science+bootcamp&restrict_sr=on

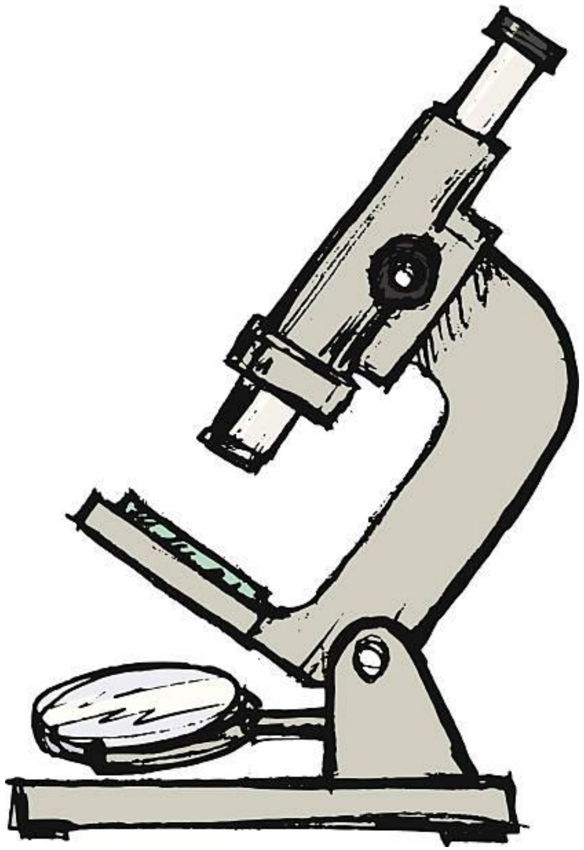
The Tools of the Trade

- Scrapy

A general framework you can install on python.

- Latent Dirichlet Allocation(NLP)
- Unsupervised Learning Algorithm

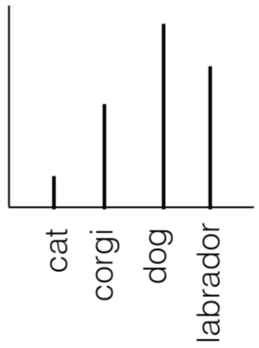
- Probabilistic Model
- The algorithm discovers patterns in the text based upon variation.



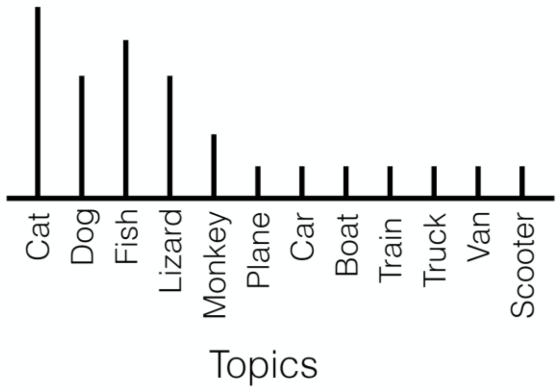
LDA continued

Topics (Z)

- A topic is a probability distribution over words

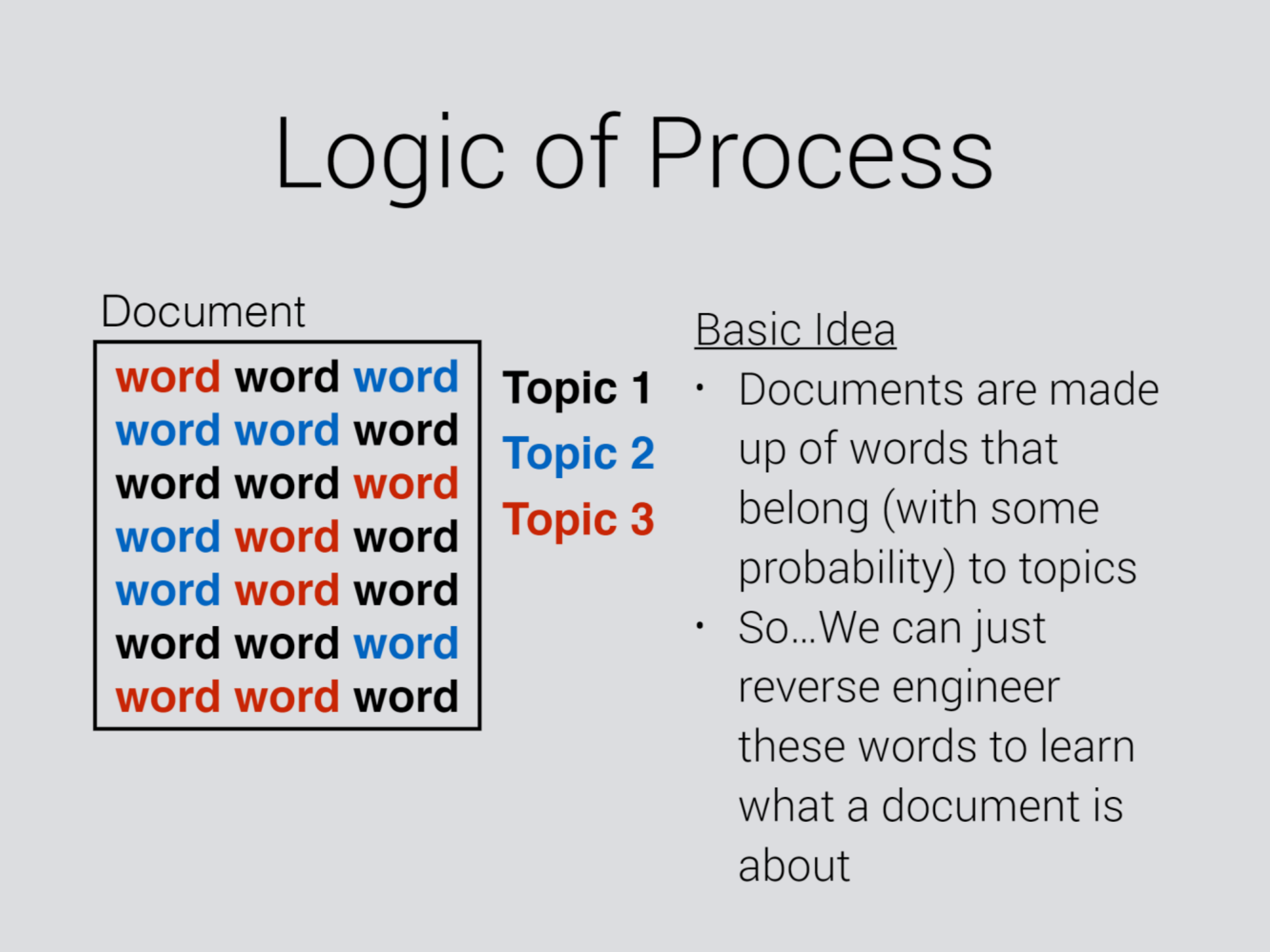


Topic Distribution for a Document



A document can be described by a recipe of topics and “how much” of each topic it contains

LDA continued



Logic of Process

Document

word word word
word word word
word word word
word word word
word word word
word word word
word word word

Topic 1
Topic 2
Topic 3

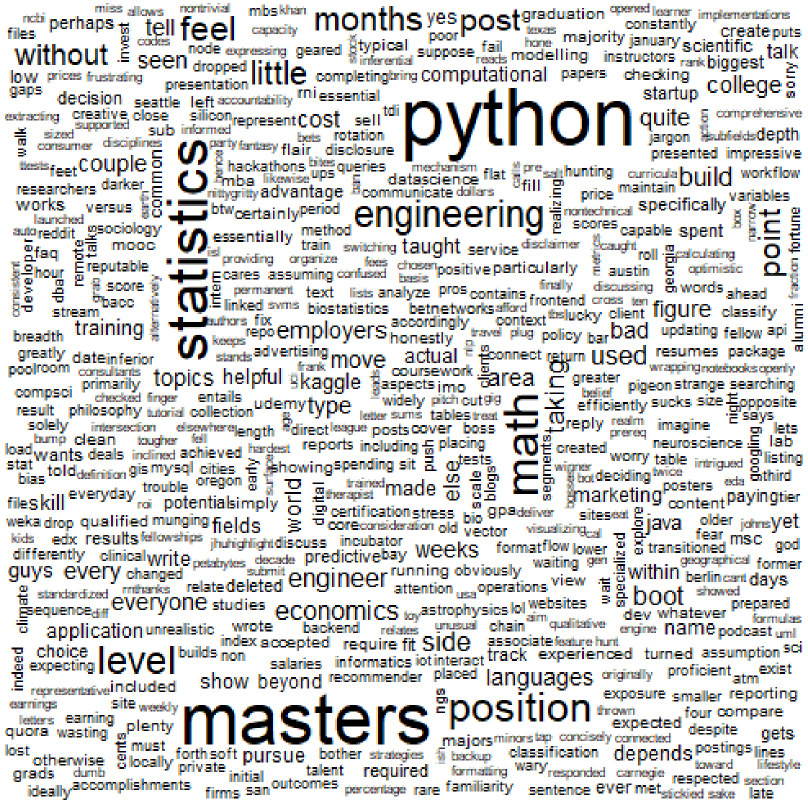
Basic Idea

- Documents are made up of words that belong (with some probability) to topics
- So…We can just reverse engineer these words to learn what a document is about

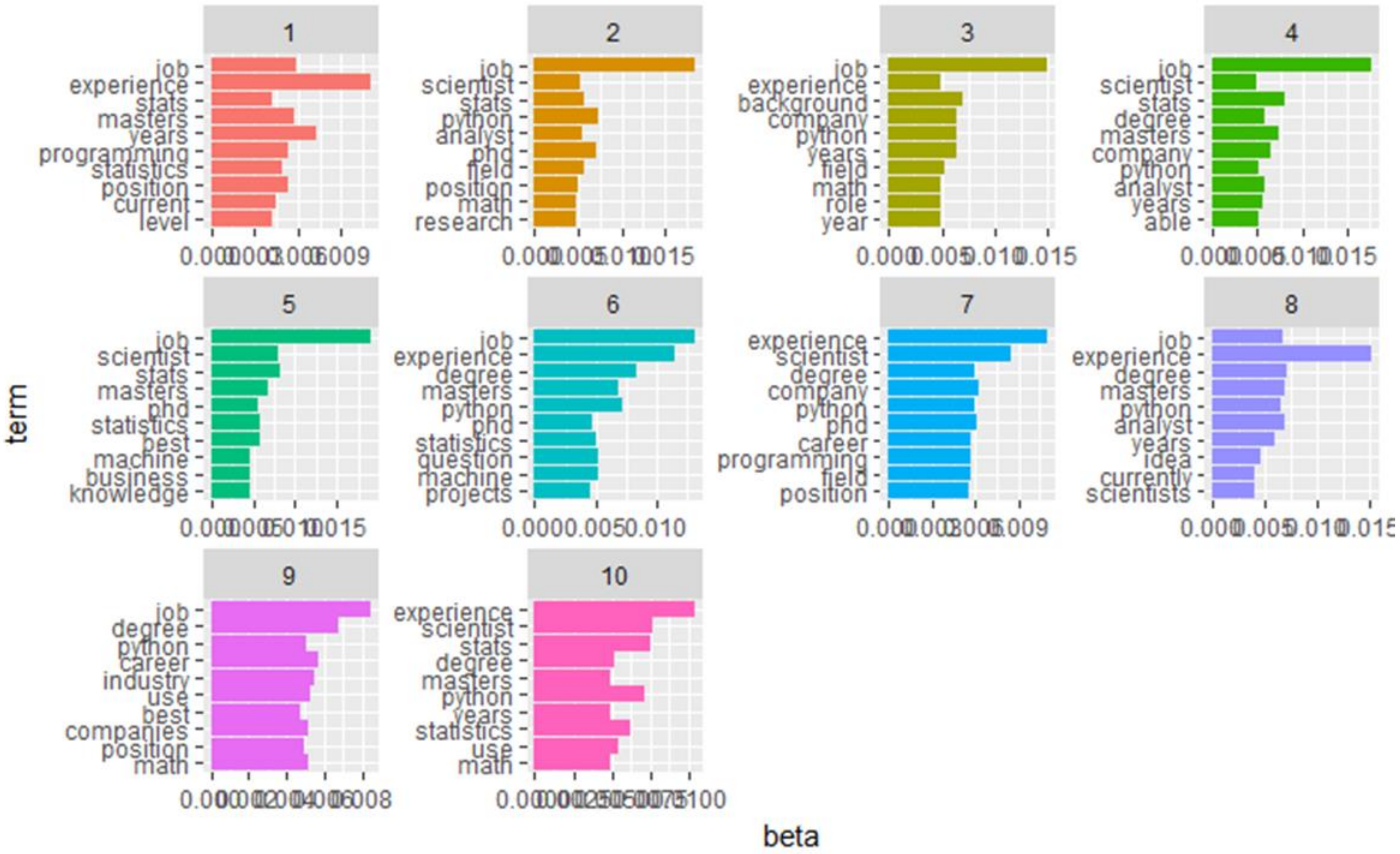
Word Cloud

122,336 words scraped

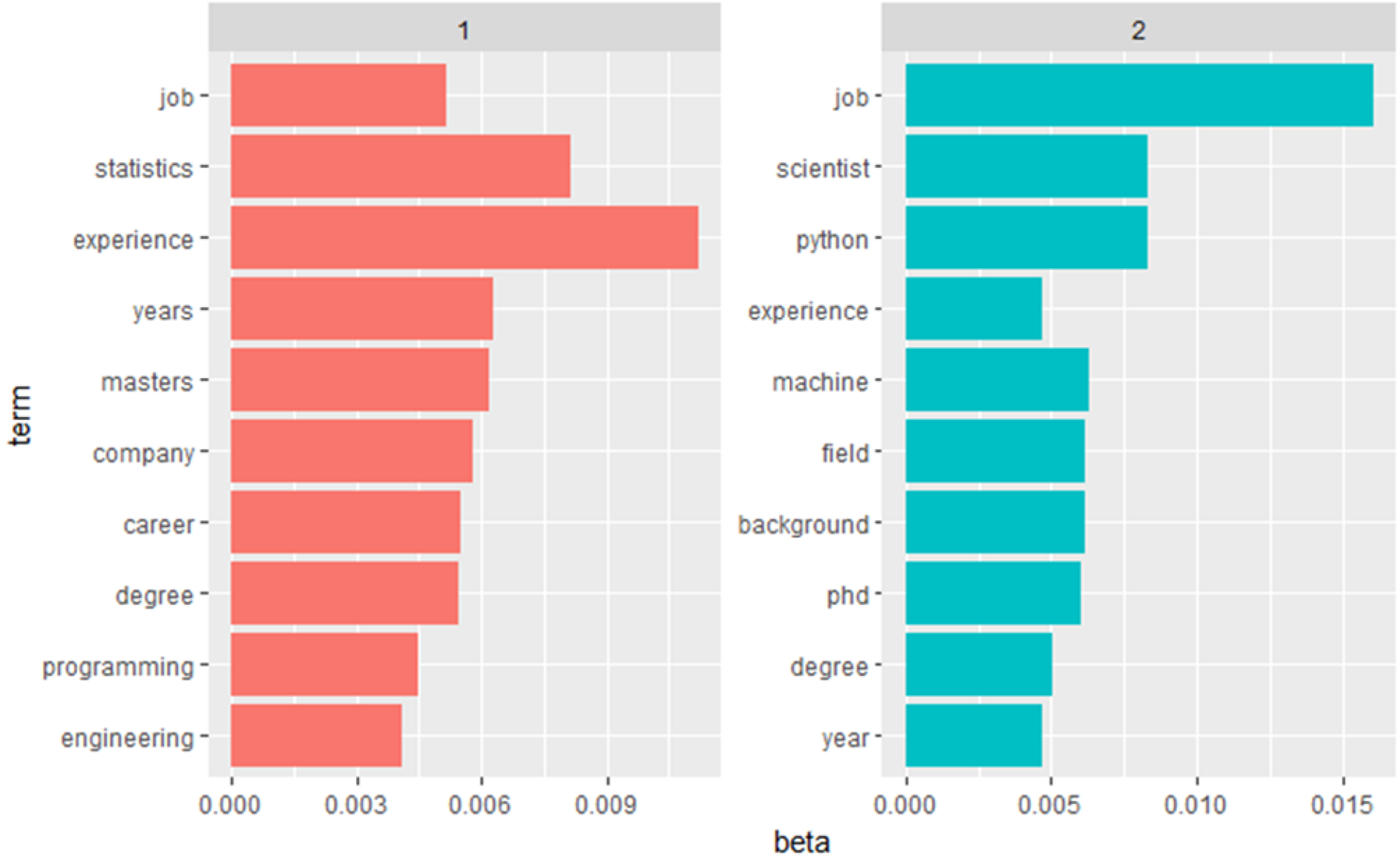
Total of 5 page lists



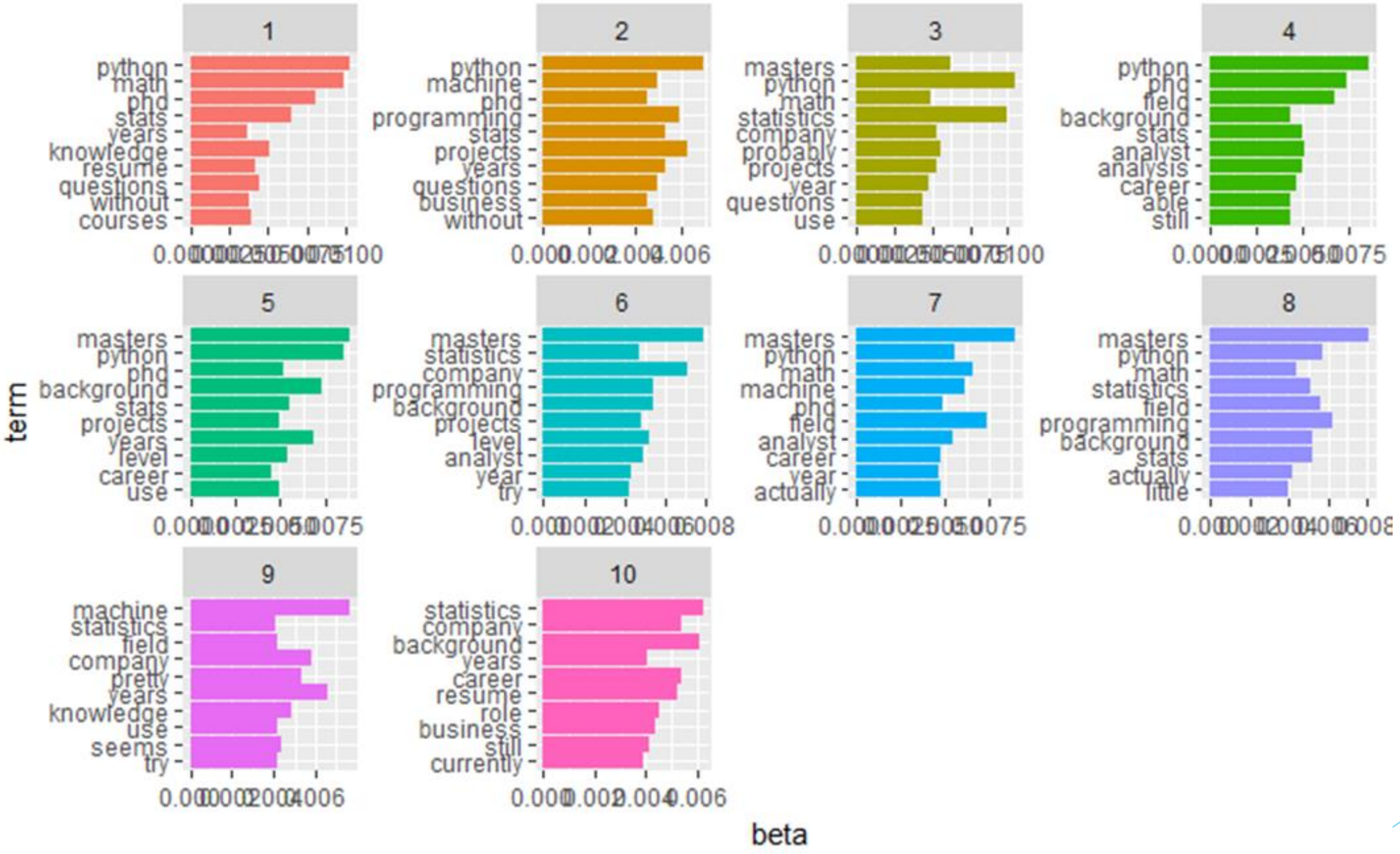
Results


1
job
experience
stats
masters
years
programming
statistics
position
current
level
0.000 0.003 0.006 0.009
2
job
scientist
stats
python
analyst
phd
field
position
math
research
0.000 0.005 0.010 0.015
3
job
experience
background
company
python
years
field
math
role
year
0.000 0.005 0.010 0.015
4
job
scientist
stats
degree
masters
company
python
analyst
years
able
0.000 0.005 0.010 0.015
5
job
scientist
stats
masters
phd
statistics
best
machine
business
knowledge
0.000 0.005 0.010 0.015
6
job
experience
degree
masters
python
phd
statistics
question
machine
projects
0.000 0.005 0.010
7
experience
scientist
degree
company
python
phd
career
programming
field
position
0.000 0.003 0.006 0.009
8
job
experience
degree
masters
python
analyst
years
idea
currently
scientists
0.000 0.005 0.010 0.015
9
job
degree
python
career
industry
use
best
companies
position
math
0.000 0.002 0.004 0.006 0.008
10
experience
scientist
stats
degree
masters
python
years
statistics
use
math
0.0000 0.0025 0.0050 0.0075 0.0100
term
beta

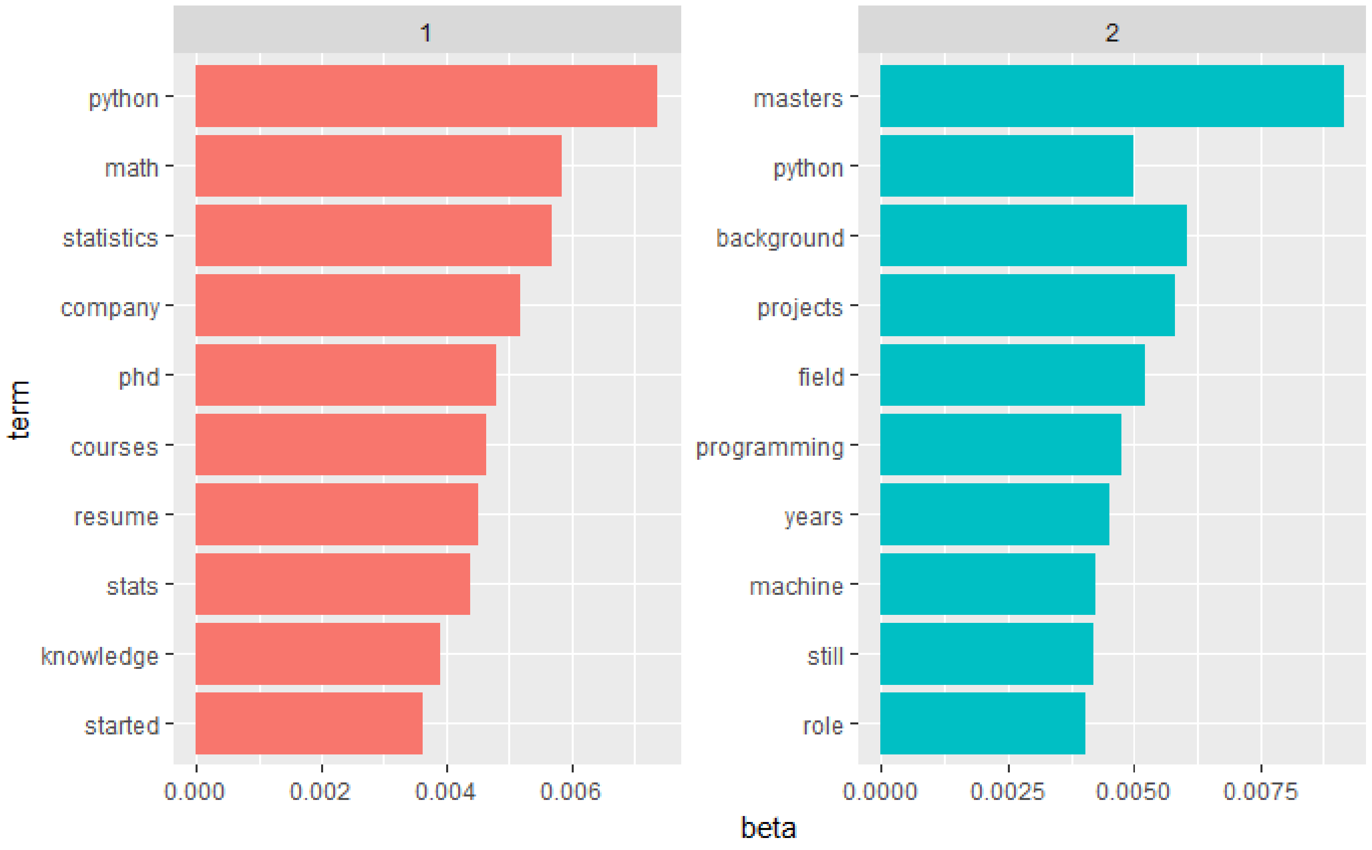
Continued


1
2
job
statistics
experience
years
masters
company
career
degree
programming
engineering
job
scientist
python
experience
machine
field
background
phd
degree
year
term
0.000
0.003
0.006
0.009
0.000
0.005
0.010
0.015
beta

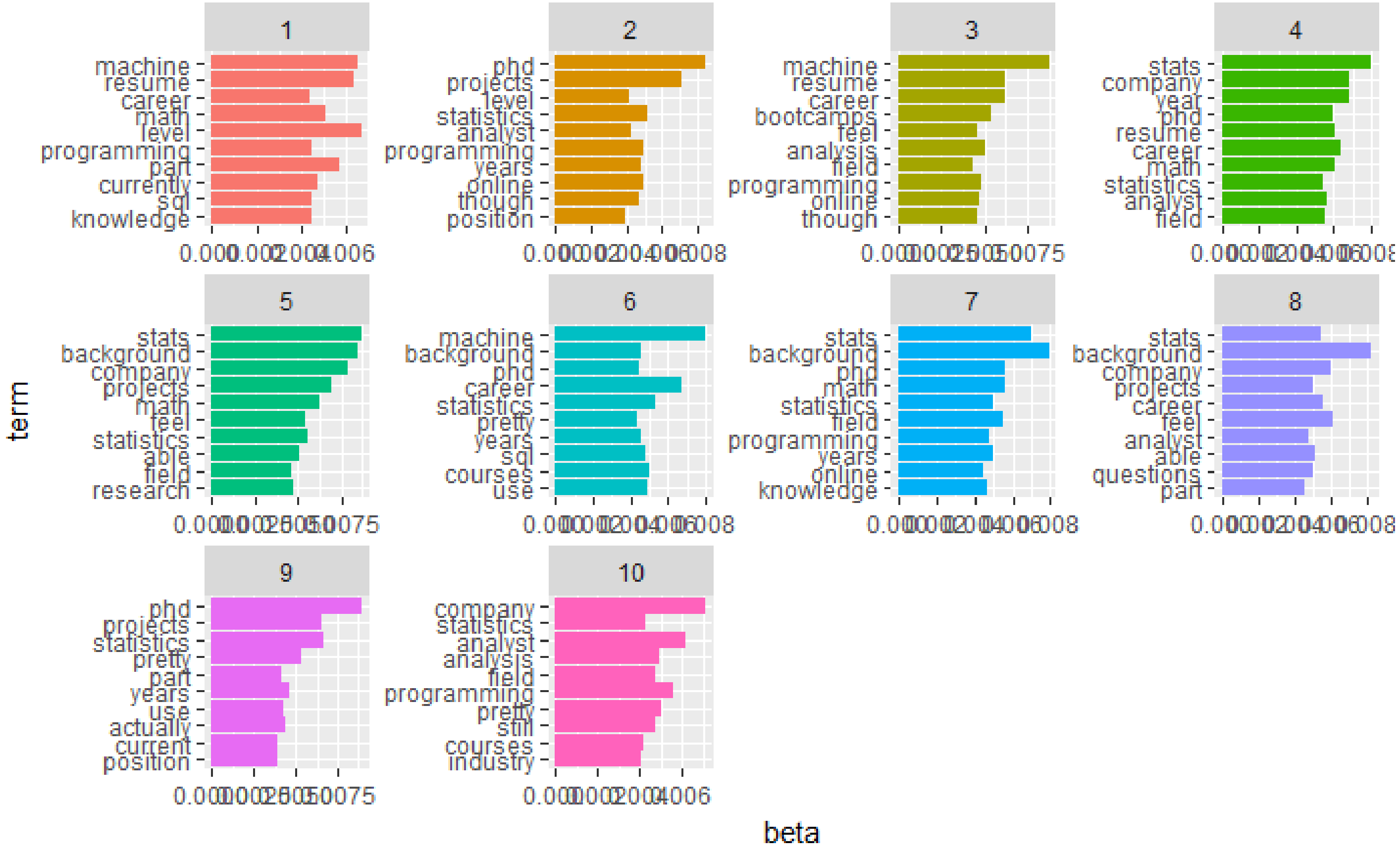
Continued


1
python
math
phd
stats
years
knowledge
resume
questions
without
courses
0.000 0.0025 0.0050 0.0075 0.0100
2
python
machine
phd
programming
stats
projects
years
questions
business
without
0.000 0.002 0.004 0.006
3
masters
python
math
statistics
company
probably
projects
year
questions
use
0.0000 0.0025 0.0050 0.0075 0.0100
4
python
phd
field
background
stats
analyst
analysis
career
able
still
0.000 0.0025 0.0050 0.0075
5
masters
python
phd
background
stats
projects
years
level
career
use
0.000 0.0025 0.0050 0.0075
6
masters
statistics
company
programming
background
projects
level
analyst
year
try
0.000 0.002 0.004 0.006 0.008
7
masters
python
math
machine
phd
field
analyst
career
year
actually
0.000 0.0025 0.0050 0.0075
8
masters
python
math
statistics
field
programming
background
stats
actually
little
0.000 0.002 0.004 0.006 0.008
9
machine
statistics
field
company
pretty
years
knowledge
use
seems
try
0.000 0.002 0.004 0.006
10
statistics
company
background
years
career
resume
role
business
still
currently
0.000 0.002 0.004 0.006
term
beta

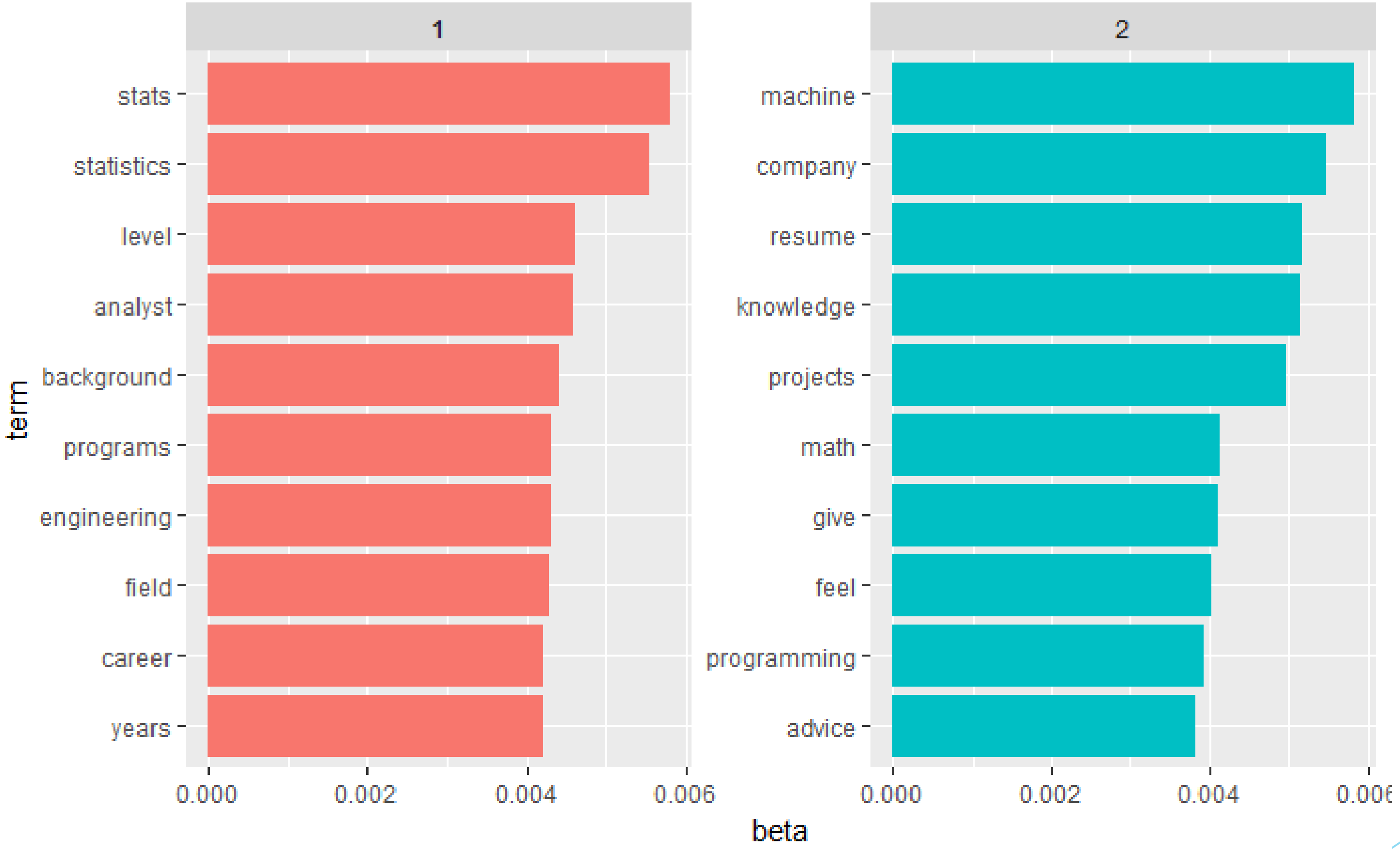
Continued


1
python
math
statistics
company
phd
courses
resume
stats
knowledge
started
term
0.000
0.002
0.004
0.006
2
masters
python
background
projects
field
programming
years
machine
still
role
0.0000
0.0025
0.0050
0.0075
beta

Continued


1
machine
resume
career
math
level
programming
part
currently
sql
knowledge
0.000 0.002 0.004 0.006
2
phd
projects
level
statistics
analyst
programming
years
online
though
position
0.000 0.002 0.004 0.006 0.008
3
machine
resume
career
bootcamps
feel
analysis
field
programming
online
though
0.0000 0.0025 0.0050 0.0075
4
stats
company
year
phd
resume
career
math
statistics
analyst
field
0.000 0.002 0.004 0.006 0.008
5
stats
background
company
projects
math
feel
statistics
able
field
research
0.0000 0.0025 0.0050 0.0075
6
machine
background
phd
career
statistics
pretty
years
sql
courses
use
0.000 0.002 0.004 0.006 0.008
7
stats
background
phd
math
statistics
field
programming
years
online
knowledge
0.000 0.002 0.004 0.006 0.008
8
stats
background
company
projects
career
feel
analyst
able
questions
part
0.000 0.002 0.004 0.006 0.008
9
phd
projects
statistics
pretty
part
years
use
actually
current
position
0.0000 0.0025 0.0050 0.0075
10
company
statistics
analyst
analysis
field
programming
pretty
still
courses
industry
0.000 0.002 0.004 0.006
term
beta

Continued


1
2
stats
statistics
level
analyst
background
programs
engineering
field
career
years
machine
company
resume
knowledge
projects
math
give
feel
programming
advice
term
0.000
0.002
0.004
0.006
0.000
0.002
0.004
0.006
beta

Conclusion: Surprise!

- Major Topics of interest amongst reddit users: Job, Experience
- Python and masters and statistics seem to be other major topics of interest
- Limitations: number of words (algorithm clustering the same terms under different topics). Generalizability (sample may not reflect the interests of people interested in data science boot camps)

The future

- Increase power of the topic modeling by adding more text and test even more numbers of topics.

- Scrape reviews based other discussion based sites like Quora and Hacker News

  And see how they relate to each other

- Compare to other programming/bootcamps

- Focus marketing strategies around related topics of interest

Keep your marketing focused on helping people find a job and host sessions python