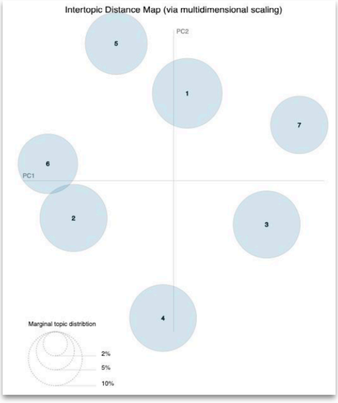
IV. Methodologies and Finding - Q1

*Randomly select **10%** of the 08-16 observations (126,422 records in total)*

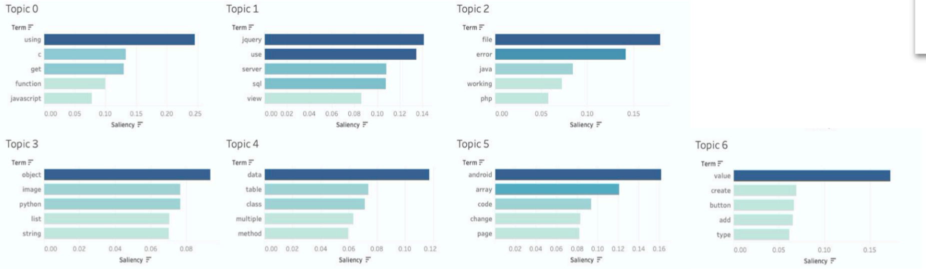
Step 1: Deduplicate by SimHash

Step 2: Train LDA model

- **Evaluation metrics**: adjust number of topics according to Vis to get the optimal result with no overlapping
- Key hyperparameters:
  no_below=1000 no_above=0.9 num_topic = 7



Step 3: Find the top 5 terms for each topic & Visualization



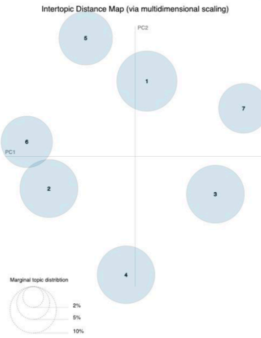
IV. Methodologies and Finding - Q1

*Randomly select **10%** of the 08-16 observations (126,422 records in total)*

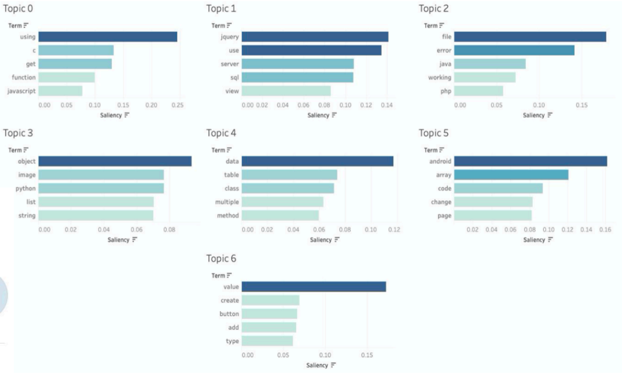
Step 1: Deduplicate by SimHash

Step 2: Train LDA model

- Evaluation metrics: adjust number of topics according to Vis to get the optimal result with no overlapping
- Key hyperparameters:
  no_below=1000
  no_above=0.9
  num_topic = 7



Step 3: Find the top 5 terms for each topic & Visualization



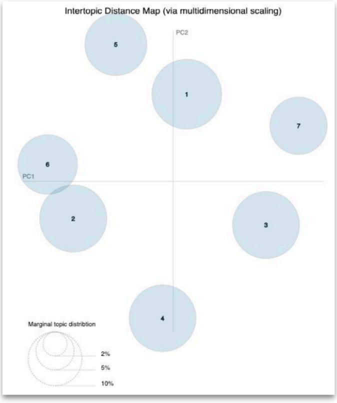
IV. Methodologies and Finding - Q1

*Randomly select **10%** of the 08-16 observations (126,422 records in total)*

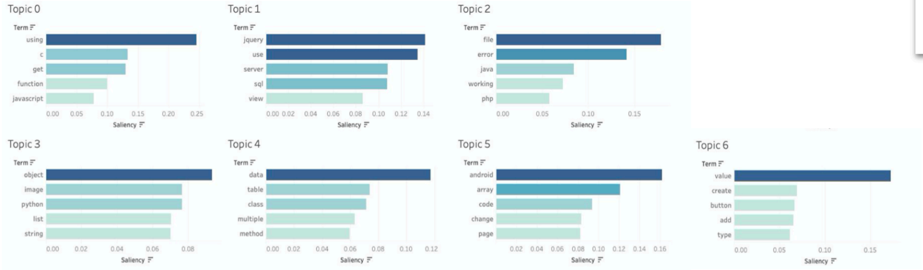
Step 1: Deduplicate by SimHash

Step 2: Train LDA model

- **Evaluation metrics**: adjust number of topics according to Vis to get the optimal result with no overlapping
- Key hyperparameters:
  no_below=1000   no_above=0.9   num_topic = 7



Step 3: Find the top 5 terms for each topic & Visualization

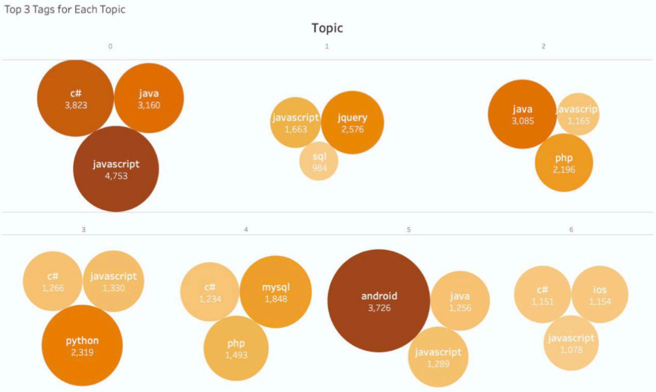


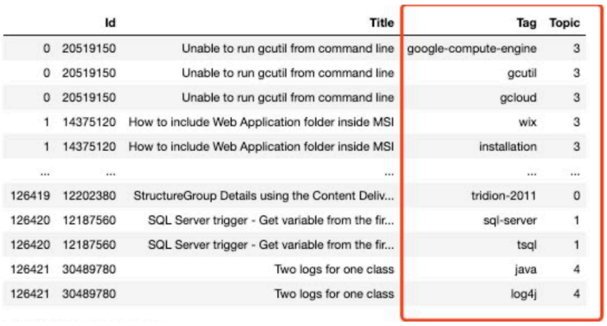
IV. Methodologies and Findings - Q1

Step 4: Join the Question file with Tags file on Id

Step 5: See which question falls into which topic group
Add topic group number to the dataframe

Step 6: Return the top 3 tag with the most counts for each topic group, and visualize the result using Tableau



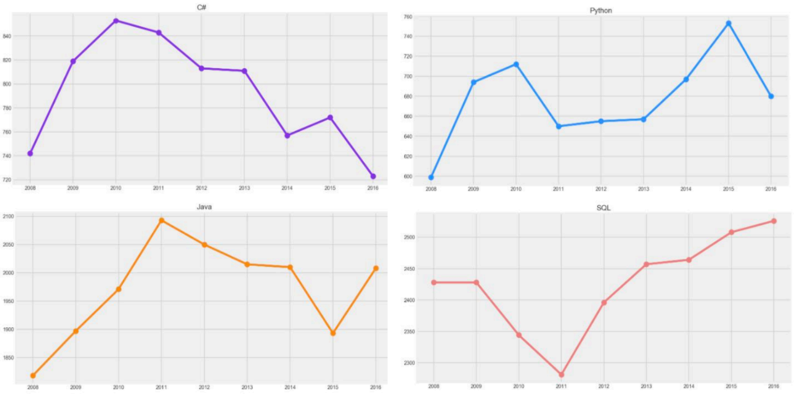


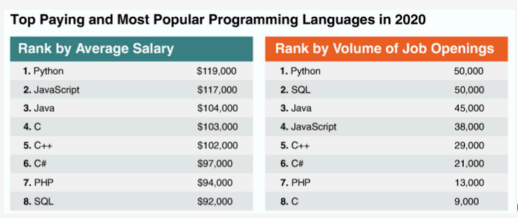
	Id	Title	Tag	Topic
0	20519150	Unable to run gcutil from command line	google-compute-engine	3
0	20519150	Unable to run gcutil from command line	gcutil	3
0	20519150	Unable to run gcutil from command line	gcloud	3
1	14375120	How to include Web Application folder inside MSI	wix	3
1	14375120	How to include Web Application folder inside MSI	installation	3
...	...	...	...	...
126419	12202380	StructureGroup Details using the Content Deliv...	tridion-2011	0
126420	12187560	SQL Server trigger - Get variable from the fir...	sql-server	1
126420	12187560	SQL Server trigger - Get variable from the fir...	tsql	1
126421	30489780	Two logs for one class	java	4
126421	30489780	Two logs for one class	log4j	4

From the result, we can see **Javascript, c#, android, python, mysql, java, php** are the dominant tags for the 7 topics of 2008-2016 stack overflow questions.

IV. Methodologies and Findings - Q2

- From the result, we can find that **Python, Java and SQL** have a growing popularity with some fluctuations, while **C#** has a significant declining popularity after 2010.
- **Evaluation metrics:** manually annotate 400 titles to check if the title was in the right topic.
- 341/400 correction in the annotation which is 85.2% accuracy rate for our model.





Top Paying and Most Popular Programming Languages in 2020

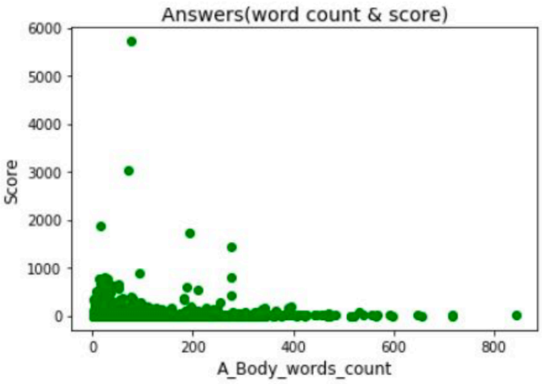
Rank by Average Salary		Rank by Volume of Job Openings	
1. Python	$119,000	**1.** Python	50,000
2. JavaScript	$117,000	**2.** SQL	50,000
3. Java	$104,000	**3.** Java	45,000
4. C	$103,000	**4.** JavaScript	38,000
5. C++	$102,000	**5.** C++	29,000
6. C#	$97,000	**6.** C#	21,000
7. PHP	$94,000	**7.** PHP	13,000
8. SQL	$92,000	**8.** C	9,000

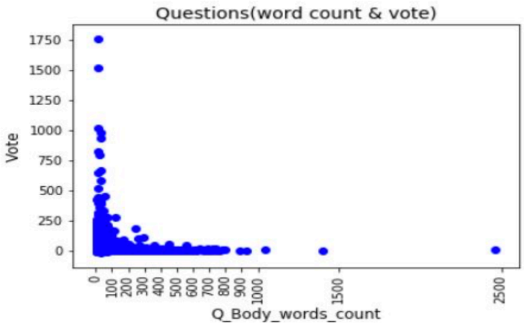
Reference:https://www.codeplatoon.org/best-paying-most-in-demand-programming-languages-2020/?gclid=Cj0KCQjwvb75BRD1ARIsAP6LcqvTgOFMAn7el6_D466eT55njJ15b3k2Sr1XNN3VsAEQqRy7rGrQb2kaArc-EALw_wcB

IV. Methodologies and Findings - Q3

Evaluation metrics: pearson correlation

- ➔ Questions (word count & vote)
  - ◆ coefficient: -0.456
  - ◆ p-value: 7.97018002483886e-20
- ➔ Answers (word count & score)
  - ◆ coefficient: 0.402
  - ◆ p-value: 2.143617417703556e-37





Advice:

- ❏ Question word counts < 200
- ❏ Answer word count < 400