LDA Output

Array of the topics, with each tuple consisting of the topic number and the relevant words.

[(0,
'0.241*"data" + 0.066*"management" + 0.065*"quality" + 0.057*"level" + 0.057*"experience" + 0.057*"unknown" + 0.043*"engineering" + 0.035*"etl" + 0.025*"governance" + 0.025*"agile"'),
(1,
'0.103*"excel" + 0.066*"finance" + 0.064*"entrylevel" + 0.057*"python" + 0.050*"seniorlevel" + 0.047*"dataanalyst" + 0.046*"power" + 0.046*"bi" + 0.045*"sql" + 0.042*"research"'),
(2,
'0.144*"learning" + 0.107*"machine" + 0.079*"engineering" + 0.067*"seniorlevel" + 0.042*"computer" + 0.036*"deep" + 0.035*"science" + 0.024*"entrylevel" + 0.022*"mathematics" + 0.021*"python"'),
(3,
'0.249*"data" + 0.072*"analysis" + 0.060*"computer" + 0.060*"science" + 0.060*"seniorlevel" + 0.055*"business" + 0.050*"analytics" + 0.047*"intelligence" + 0.035*"big" + 0.032*"visualization"'),
(4,
'0.146*"computer" + 0.113*"science" + 0.083*"seniorlevel" + 0.049*"engineering" + 0.042*"architecture" + 0.036*"vision" + 0.033*"aws" + 0.031*"classification" + 0.027*"deep" + 0.026*"learning"'),
(5,
'0.118*"aws" + 0.088*"azure" + 0.078*"data" + 0.078*"seniorlevel" + 0.066*"architecture" + 0.050*"agile" + 0.046*"big" + 0.044*"airflow" + 0.039*"apis" + 0.035*"pipelines"')]

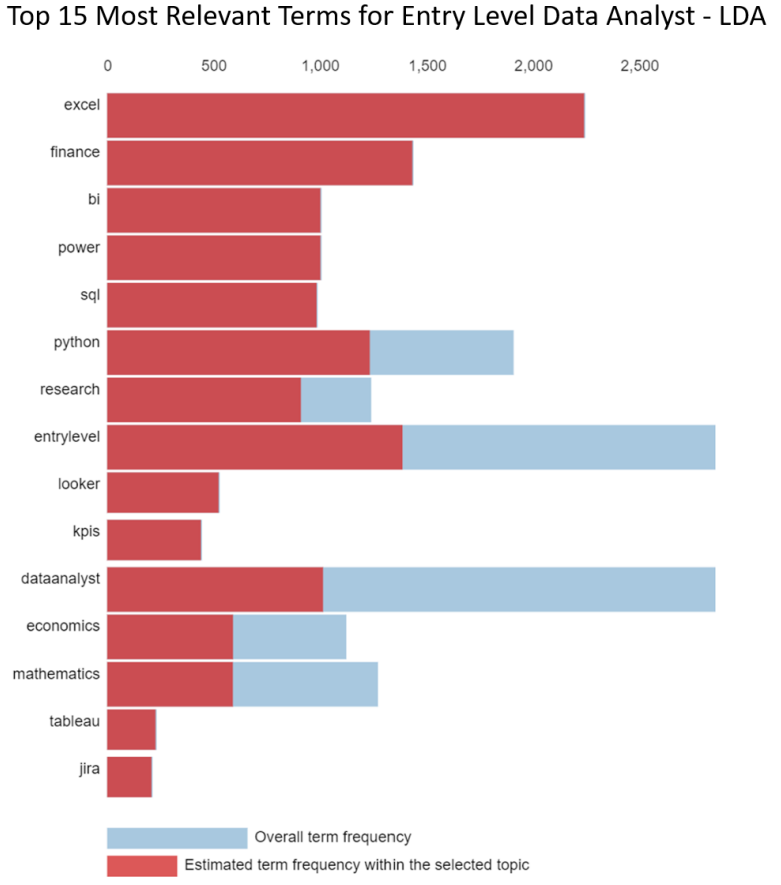


Image 2: Example of a topic generated by LDA. The image displays highly impactful terms to the topic, with importance on frequency to other terms within the same topic taking prevalence over frequency of the term across all documents.

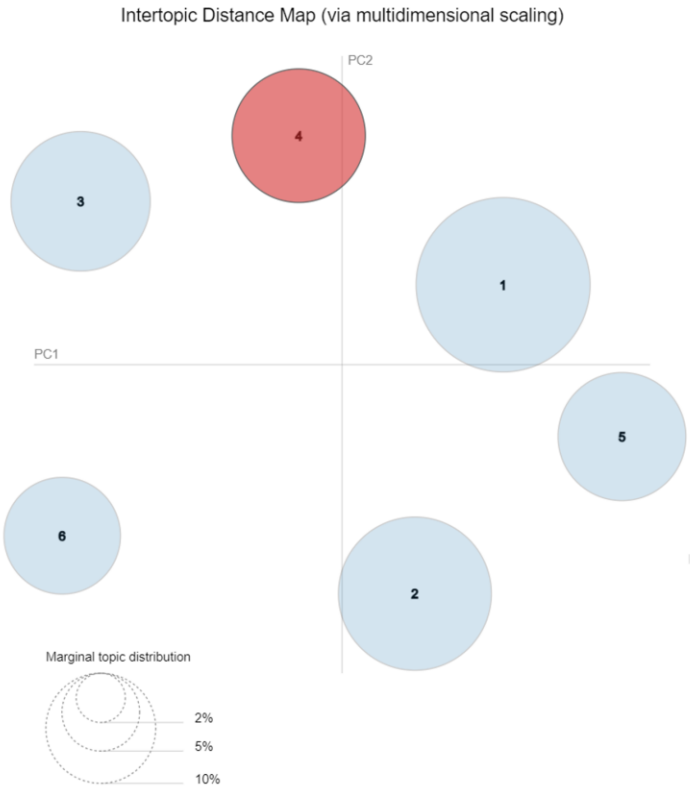


Image 3: Intertopic Distance between topics generated by LDA, generated through the pyLDAvis package.

Top2Vec

Dictionary of the top 10 words in each topic and their cosine similarity score to the centroid of the topic.

Top 10 words in topic 0:
{'level': 0.71737707, 'unknown': 0.70569026, 'data': 0.69349885, 'analysis': 0.67938375, 'consulting': 0.6535479, 'intelligence': 0.6422585, 'management': 0.64147604, 'banking': 0.6413659, 'experience': 0.6320977, 'blockchain': 0.6221915}

Top 10 words in topic 1:
{'autonomous': 0.78548694, 'chatgpt': 0.76288795, 'inference': 0.7410071, 'deep': 0.73515177, 'machine': 0.72089016, 'gpt': 0.69740826, 'design': 0.695372, 'open': 0.6947685, 'modeling': 0.6835383, 'ci': 0.6682361}

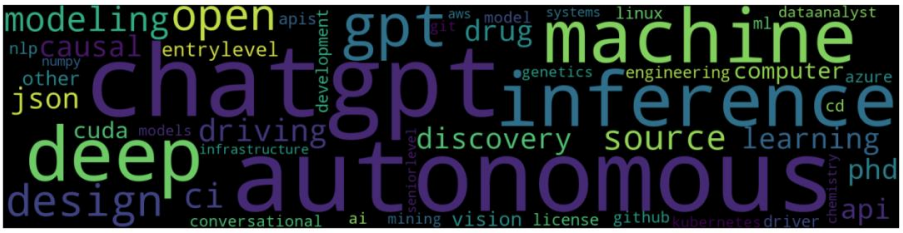
Top 10 words in topic 2:
{'unknown': 0.16803727, 'warehouse': 0.16656809, 'analysis': 0.16238183, 'snowflake': 0.16124186, 'gcp': 0.16022336, 'fintech': 0.14336355, 'google': 0.14086422, 'industrial': 0.12134726, 'clustering': 0.11226115, 'risk': 0.10709187}

Top 10 words in topic 3:
{'rdbms': 0.13984104, 'ci': 0.13982026, 'classification': 0.1256078, 'ai': 0.12371657, 'phd': 0.12214897, 'content': 0.11937801, 'quality': 0.118576616, 'bigtable': 0.11245971, 'github': 0.111968905, 'tableau': 0.10766883}

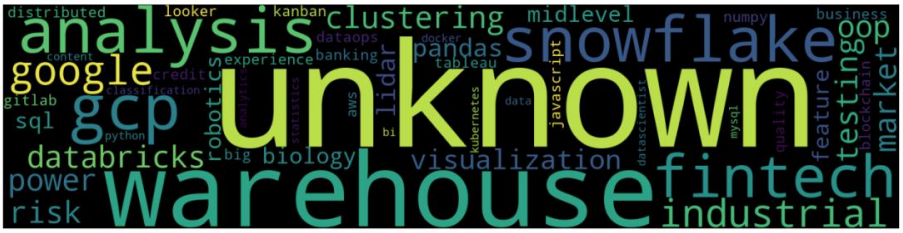
Top 10 words in topic 4:
{'etl': 0.1527841, 'physics': 0.10270392, 'sas': 0.10232733, 'rdbms': 0.09125591, 'conversational': 0.08771415, 'research': 0.08271344, 'statistics': 0.07842212, 'driver': 0.07701689, 'infrastructure': 0.06821172, 'models': 0.06695615}

Topic 0

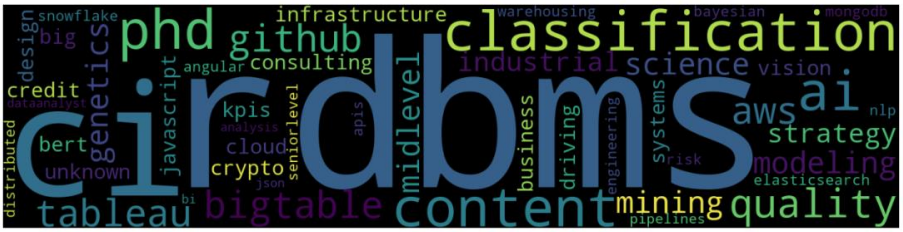
Topic 1



Topic 2



Topic 3



Topic 4

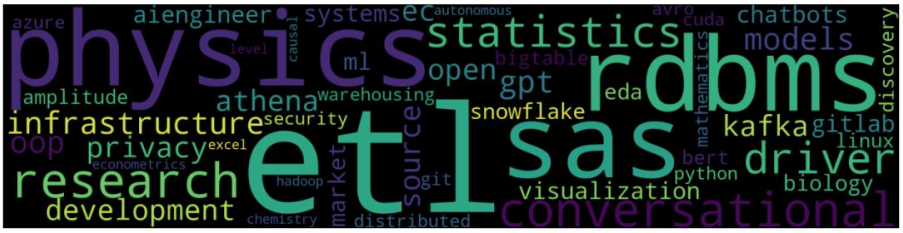


Image 4: Word clouds where the size represents the cosine similarity of the words to the centroid of their topic.

BERTopic

-1	['data', 'seniorlevel', 'learning', 'engineering', 'other', 'computer', 'science', 'machine', 'architecture', 'aws']
0	['data', 'aws', 'seniorlevel', 'azure', 'architecture', 'agile', 'big', 'airflow', 'other', 'computer']
1	['excel', 'entrylevel', 'intelligence', 'business', 'python', 'bi', 'power', 'sql', 'analysis', 'data']
2	['learning', 'machine', 'deep', 'computer', 'science', 'engineering', 'seniorlevel', 'vision', 'other', 'chatbots']
3	['blockchain', 'crypto', 'banking', 'seniorlevel', 'science', 'computer', 'airflow', 'other', 'aws', 'finance']
4	['privacy', 'research', 'other', 'machine', 'seniorlevel', 'engineering', 'learning', 'market', 'excel', 'python']
5	['docker', 'deep', 'vision', 'aws', 'computer', 'learning', 'engineering', 'cd', 'ci', 'git']
6	['causal', 'inference', 'economics', 'machine', 'learning', 'engineering', 'seniorlevel', 'datascientist', 'science', 'data']
7	['robotics', 'ecommerce', 'python', 'seniorlevel', 'engineering', 'php', 'cad', 'statistics', 'linux', 'other']
8	['angular', 'testing', 'javascript', 'apis', 'react', 'bigquery', 'typescript', 'seniorlevel', 'architecture', 'engineering']

Image 5: Visualization of the distance between topics.