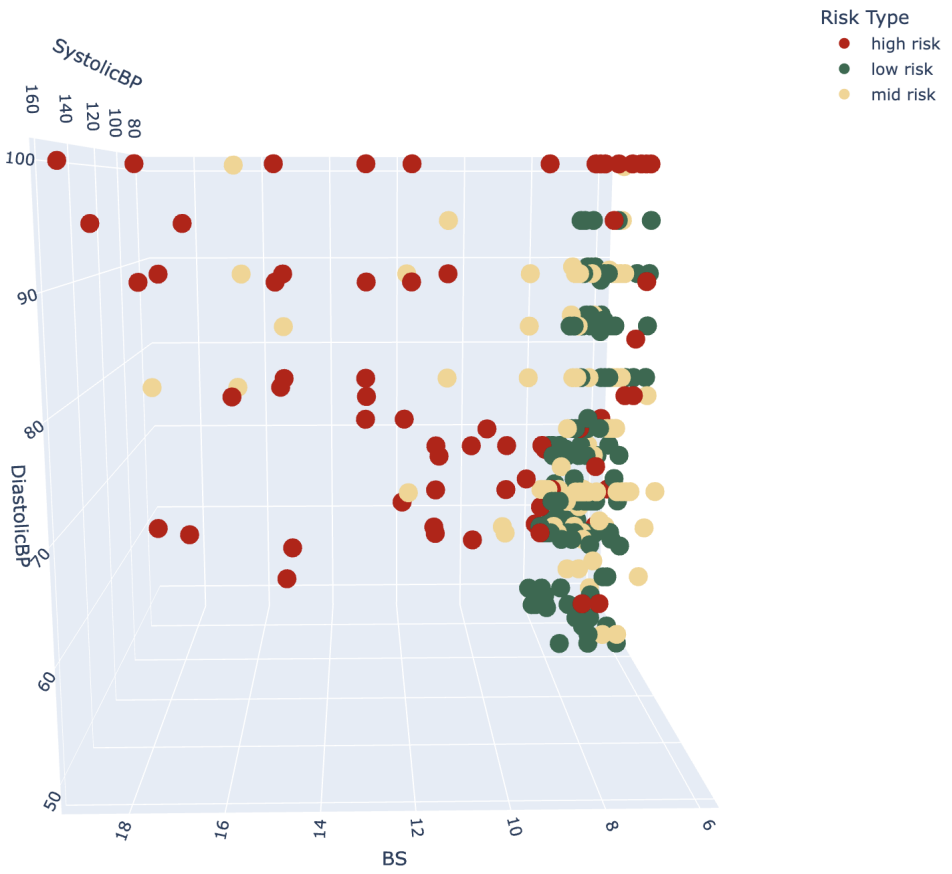
TBD

TBD

TBD

Jonas Michel

Student ID: 24238749



TBD

1 Question 1 (10 Marks)

Given a query that a user submits to an IR system and the top N documents that are returned as relevant by the system, devise an approach (high-level algorithmic steps will suffice)

to suggest query terms to add to the query. Typically, we wish to give a large range of suggestions to the users capturing potential intended query needs, i.e., high diversity of terms that may capture the intended query context/content.

Ziel: query expanden können - which makes the querey better the terms should be relevant (able to expand the query in a meaningful way - find the correct topic), diverse (cover different topics)

we should design a heuristic allowing that

Ansatz: clusters, there are papers on that

1. query with q_n terms, for the whole vocuabulary we get the most similar terms (maybe 1000 t_n) 2. cluster the t_n terms into k clusters - then we get the most central term from every cluster

no terms that were in the query

how to get the most describtive terms for the query

when - runtime vs offline

there are multiple expansion methods out here: like sysnonm, related term, contextual,

recall vs precision tradeoff what we trying to solve and how we are acutal doing it

depending what a system we are designing (incooperate user data, e.g. Galway, Ireland, Europe)

user can give a temperature - to select the cluster

2 Section TBD.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Appendix A

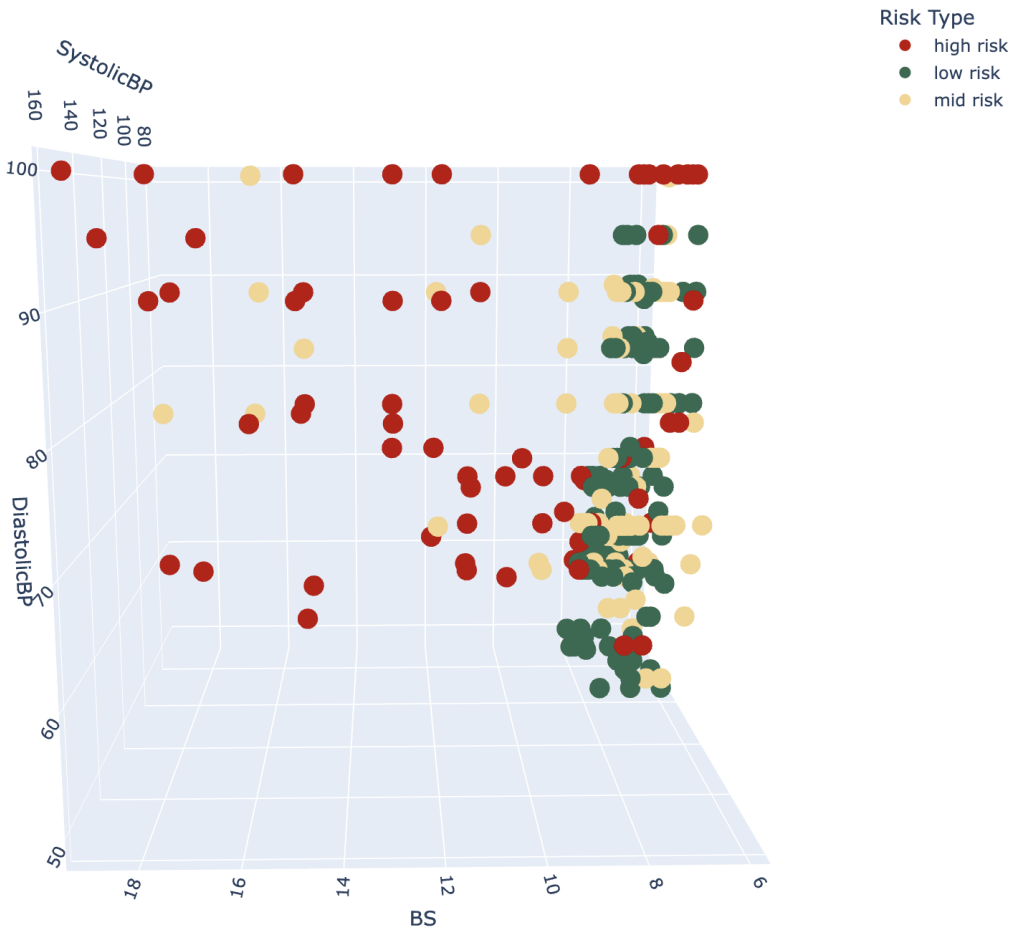


Figure 1: The actual caption

Appendix B

	Low Risk	**Mid Risk**	**High Risk**
Total Number of Entries	404	336	272
Percentage	40%	33%	27%

Table 1: Class Distribution within the Dataset

	Decision Tree			**KNN**		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Low Risk	0.90	0.75	0.82	0.67	0.81	0.74
Mid Risk	0.76	0.84	0.80	0.74	0.51	0.60
High Risk	0.84	0.89	0.87	0.77	0.86	0.81
Weighted	0.83	0.83	0.83	0.72	0.72	0.72

Table 2: Numerical Results