Aluno: Pedro Henrique Silva Santana

Matrícula: 12011BSI218

ES4 - Modelo Probabilístico

D1 – {logan e ororo são x-men}

D2 – {stark, parker e logan já foram vingadores parker gostaria de ser novamente}

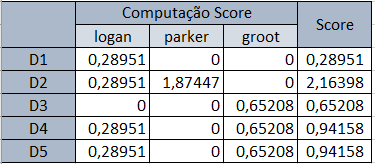
D3 – {ororo e stark não são guardiões como groot e rocket}

D4 – {eu sou groot, logan, todos somos groot e groot precisa de ajuda}

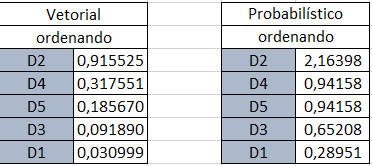
D5 - {logan e rocket formariam uma dupla melhor do que logan e groot}

Assuma que o vocabulário dessa coleção seja formado pelos seguintes termos de indexação: V = {logan, ororo, stark, parker, groot, rocket}.

a) Considere a seguinte consulta q1 = {logan parker groot}. Calcule o grau de similaridade dos documentos da coleção para a consulta q1 usando o modelo probabilístico (considere R=ri=0).



b) Compare os rankings obtidos na letra a) com os rankings obtidos no ES-3 (modelo vetorial). Quais as semelhanças/diferenças? Como você poderia explicar tais semelhanças/diferenças?



Os dois rankings possuem semelhança na ordenação dos documentos devido ao fato de não ser atribuído relevância para esta parte do probabilístico. São diferentes nos valores obtidos e no caso, entre d4 e d5, houve um empate, porem no vetorial esse empate não existe.

c) Suponha que a coleção possua 100.000 documentos. Considere q2 = {stark rocket}. Qual procedimento poderia ser utilizado para encontrar estimativas para R e ri? Detalhe tal procedimento usando a consulta q2 como exemplo.

Inicialmente seria necessário fazer uma busca inicial definindo R = ri = 0, depois selecionar o top 5000 (5%) e definir os relevantes (ri) para a consulta (stark rocket). Depois disso aplicar os dados mensurados com base nessa metodologia e remover os documentos antes analisados e, assim, refazer a consulta.