

Team Information

Student 1: 12126329 Marko Pajkovski

Part 1

In order to improve the current aggregation of judgments, we need to think more deeply about the judgment process. Whenever, a user is given a document-query, he is able to choose between four different relevance options, ranging from not relevant to perfect answer. However, each user has a propensity to judge in a personalised way i.e, some users require more matched information between the query and document to judge high, while others do not. Thus, I propose a way to normalise the deviation of judgments between users, by centering judgments around each users mean. Secondly, by the assumption that all users take their time to evaluate and analyze the document-query pairs, I proposed a metric based on the overall user disagreement of each judgment, that penalises the relevancy based on the disagreement metric.

Assumption: All users take their time to evaluate and analyze the document-query pairs.

Hypothesis 1: Each user has a personalised way of evaluating the document-query pairs, thus centering.

Hypothesis 2: If there is disagreement between users the document is not as relevant as the majority/mean judgment.

$$J(d_i) = \begin{cases} \text{mean}(j_d i) - D(d_i), & \text{if mode tie exists} \\ \text{mode}(j_d i) - D(d_i), & \text{otherwise} \end{cases}$$

$$\text{Where disagreement: } D(d_i) = \frac{\sum_{j \neq k}^{judges} \frac{r_j - r_k}{2}}{||judges||}$$

Meta Judgment

Document-Query 1:

Base Judgment: 2 vs My Judgment: 1

Query: p7011

Document: ft 12 may 93 / observer : meant well spotted in the visitors book of a hotel in india . ' a wonderful stay - spoilt by the staff ' . countries : - inz india , asia . industries : - p7011 hotels and motels . types : - news general news . the financial times london page 19

Here the document contains p7011 however it is not relevant

Document-Query 2: Base Judgment: 2 vs My Judgment: 1

Query: jamul newspaper

Document: metro ; part b ; page 7 ; column 3 ; metro desk 1 sailor dead , 3 hurt in jamul car crash a 21 - year - old navy man was killed early saturday morning when he lost control of his car and was thrown from it , the california highway patrol said . his passengers , who were also ejected from the vehicle , sustained moderate to major injuries . the man was heading east on route 94 at 4 : 10 a . m . near melody road in jamul when he ran off the right shoulder of the road . he overcorrected , causing the car to flip several times . neither the driver nor the passengers were wearing seat belts . the driver was pronounced dead at the scene . the three passengers were taken to mercy and sharp hospitals .

Here the query is jamul newspaper however we do not now if the document is from a jamul newspaper, or if it is just an article for a jamul car crash

Document-Query 3: Base Judgment: 4 vs My Judgment: 2

Query: north carolina sports

Document: " " " " " " sports ; part c ; page 14 ; column 1 ; sports desk
newswire the police officer who arrested north carolina basketball player king rice on
assault charges and resisting arrest exploited the point guard ' s celebrity status , coach
dean smith said . " " " " " " i don ' t know what his deal is , " " " " " "
" " " " " " smith told the chapel hill (n . c .) newspaper wednesday . " " " " " "

""" """ i hear (the officer) used to follow j . r . (reid) around , too . """
 """ """ reid is a former north carolina basketball star playing in the nba . rice ,
 21 , will appear in district court may 31 to face charges he assaulted natalie nester ,
 his girlfriend , a member of the north carolina state women ' s team . column college
 athletics ; basketball players ; university of north carolina ; assaults - - north carolina
 """ """

Here the query refers to sports in north carolina, however this is about an arrest for a north carolina sports player.

Part 2

In this part only the classes KNRM and Conv KNRM have been implemented, which were just copied from Phd candidate Sebastian Hofstätter.

Part 3

In this part I have used the Minilm uncased squad2 from Deepset. Moreover, I created another metric that checks if the selected 'gold' answer is a subset of the model answer. The final results are the following:

Total Evaluations	52606
Gold answer subset of Model Answer	9441
Exact Match	6867
Average F1-Score	0.4368