Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Department of Information Engineering

Master's Degree in Computer Engineering

Development of an IR system for argument search

Touché Task 1: Argument Retrieval for Controversial Questions

HyperGroup / Jean Pierre Polnareff

Marco Alecci, Tommaso Baldo, Luca Martinelli, Elia Ziroldo

ACADEMIC YEAR 2020/2021

Introduction

Task 1: Supporting debates on controversial topics

Scenario:

Users search for arguments on controversial topics

Task:

Retrieve “strong” pro/con arguments on the topic

Data:

args.me corpus, a collection of documents extracted from web debate portals

Related works: overview of Touché 2020

General Strategy:

1. Base retrieval model
   - *LMDirichlet* and *BM25* most used
   - *LMDirichlet* and *DPH* most performant
2. Augmentation
   - Query expansion
   - Result expansion
3. Re-ranking
   - Argument quality
   - Sentiment analysis

Team	Retrieval	Augmentation	(Re)ranking Feature
Dread Pirate Roberts	DirichletLM/Similarity -based	Language modeling	
Weiss Schnee	DPH	Embeddings	Quality
Prince of Persia	Multiple models	Synonyms	Sentiment
The Three Mouseketeers	DirichletLM		
Swordsman (Baseline)	DirichletLM		
Thongor	BM25/DirichletLM		
Oscar François de Jarjayes	DPH/Similarity-based		Sentiment
Black Knight	TF-IDF	Cluster-based	Stance, readability
Utena Tenjou	BM25		
Arya Stark	BM25		
Don Quixote	Divergence from Randomness	Cluster-based	Quality + Similarity
Boromir	Similarity-based	Topic modeling	Author credibility
Aragorn	BM25		Premise prediction
Zorro	BM25		Quality + NER

Methodology

- Pre-processing of the documents
- BM25 and LMDirichlet
- Different strategies, first separately than merged:

  - Different weight to different fields of the document
  - Query expansion using synonyms extracted from WordNet
  - Re-ranking based on sentiment analysis on the documents

Pre-processing

- Classic Tokenizer
- LowerCaseFilter and LengthFilter
- Custom filter: MultipleCharFilter

Pre-processing: choice of stoplist

- Different stock stoplists
- Best scores with shorts stoplists
- Max score: EBSCOhost

Stock stoplists	Number of words	nDCG@5
tent1	400	0.5599
Air3z4	1298	0.5757
zettair	469	0.5790
smart	571	0.5895
terrier	733	0.5919
cook	221	0.6043
taporwave	485	0.6068
postgre	127	0.6078
nltk	153	0.6078
lexisnexis	100	0.6131
NO STOPLIST	0	0.6189
corenlp	28	0.6211
okapi	108	0.6224
ranksnl	32	0.6249
lucene_elastic	33	0.6256
ovid	39	0.6259
lingpipe	76	0.6260
EBSCOhost	**24**	**0.6265**

Pre-processing: custom stoplist

Custom stoplists	Number of words	nDCG@5
150_custom	150	0.6066
ebsco+10	34	0.6258
ebsco+20	44	0.6258
ebsco+30	54	0.6123

- **150_custom:** 150 most frequent terms in the index
- **Ebsco+x:** ebsco stoplist with respectively the 10, 20 and 30 most frequent terms in the index (not already in the stoplist)

Pre-processing: stemmers

Stem Filter	**nDCG@5**
No Stem	**0.6265**
English Minimal Stem	0.6184
Krovetz Stem	0.5747
Porter Stem	0.5401

Adding complexity to the system, the score obtained decreases, this probably due to limitations of stemmers used

Strategies

1) Different weights to fields

- Three differents fields: Body, Premises, Conclusions
- All combinations of weights tested from 0 to 1, with a step of 0.25

BM25

Body	Premises	Conclusions	nDCG@5
0.0	1.0	0.25	0.4150
0.25	1.0	0.25	0.4143
0.5	1.0	0.25	0.4032
0.5	0.75	0.25	0.4029
0.25	0.75	0.25	0.4023

LMDirichlet

Body	Premises	Conclusions	nDCG@5
0.25	1	0	0.7379
0	1	0	0.7345
0.25	0.75	0	0.7331
0.5	1	0	0.7239
0.5	0.75	0	0.7123

2) Query expansion: synoynms

- Add synonyms to query before the search
- **WordNet:** lexical database of semantic relations between words

Synonyms Weight	BM25	LMDirichlet
No synoynms	0.3938	**0.6339**
0.1	0.4113	0.6185
0.2	**0.4159**	0.5977
0.3	0.3973	0.5913
0.4	0.3898	0.5267
0.5	0.3764	0.4731
0.6	0.3596	0.4273
0.7	0.3304	0.3847
0.8	0.2931	0.3406
0.9	0.2584	0.2892
1.0	0.2253	0.2564

3) Re-Ranking: Sentiment analysis

- **Compute a value between -1 and 1 for each argument:**
  - Greater than zero: positive sentiment
  - Lower than zero: negative sentiment

- **Two approaches tried both on Conclusions and Premises:**
  - Priorities to neutral documents
  - Priorities to emotional documents

Premises

Premises	**BM25**	**LMDirichlet**
No sentiment	0.3938	**0.7345**
Neutral is better	0.0811	0.0569
Emotional is better	**0.4362**	0.6952

Conclusions

Conclusion	**BM25**	**LMDirichlet**
No sentiment	**0.3938**	**0.7345**
Neutral is better	0.0811	0.0569
Emotional is better	0.1423	0.1414

Results

Run	BM25	LMDirichlet
Base	0.3938	0.7345
Best different fields weights	**0.4698**	**0.8026**
Best query expansion with synonyms	0.4159	0.6986
Best re-ranking with sentiment analysis	0.4362	0.6952
Merging all three strategies	0.4521	0.6661

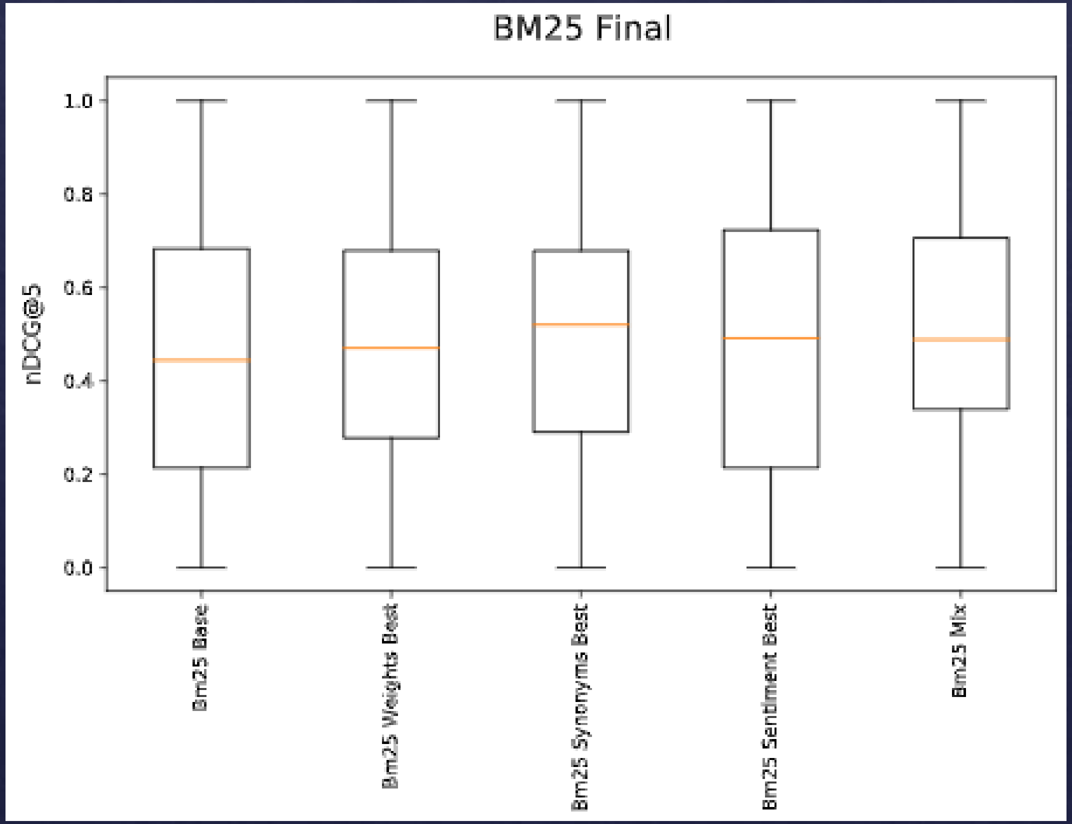
- **BM25**: all scores improved
- **LMDirichlet**: Query expansion and Re-ranking do not work well

LMDirichlet better than BM25

Statistical Analysis

BM25 Best Runs

1) Boxplots



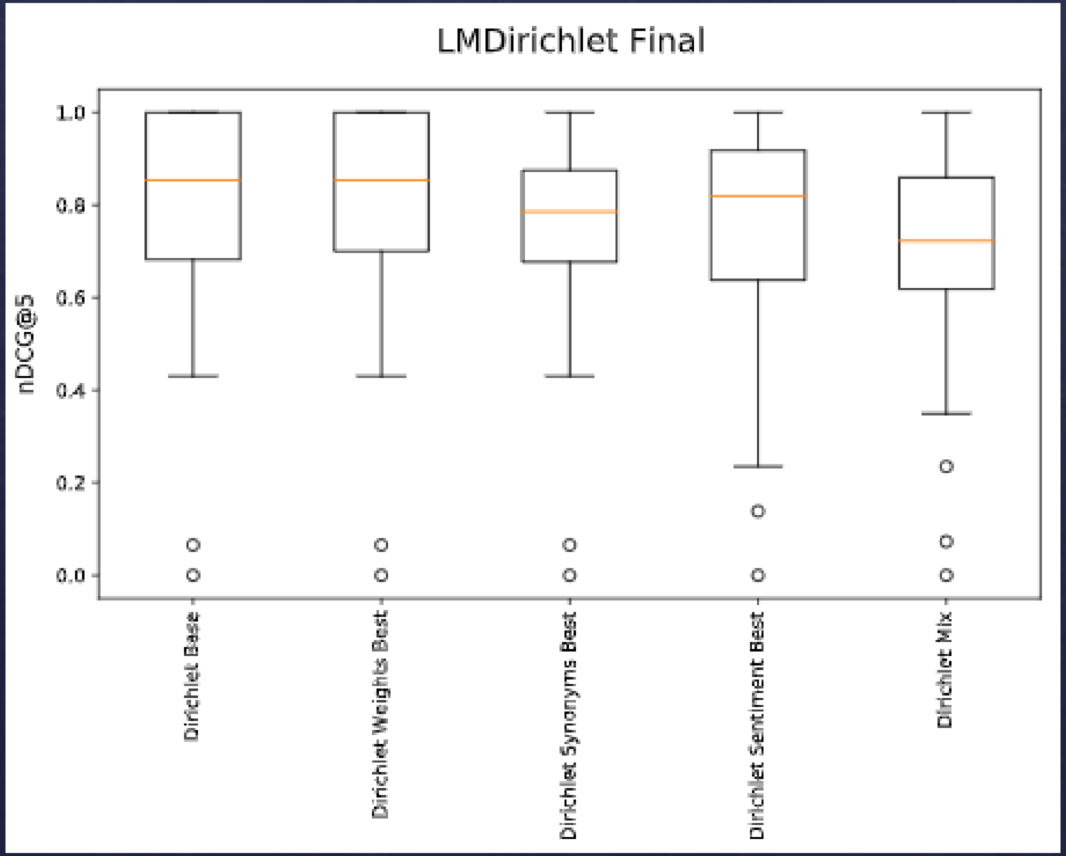
2)**ANOVA** Test:
p-value = **0.846705**

3) Multiple pairwise comparison (Tukey's HSD Test)

run1	run2	p-value
BM25 Base	BM25 Weights Best	0.9
BM25 Base	BM25 Synonyms Best	0.9
BM25 Base	BM25 Sentiment Best	0.9
BM25 Base	BM25 Mix	0.749016
BM25 Weights Best	BM25 Synonyms Best	0.9
BM25 Weights Best	BM25 Sentiment Best	0.9
BM25 Weights Best	BM25 Mix	0.9
BM25 Synonyms Best	BM25 Sentiment Best	0.9
BM25 Synonyms Best	BM25 Mix	0.9
BM25 Synonyms Best	BM25 Mix	0.9

LMDirichlet Best Runs

1) Boxplots



2) **ANOVA** Test:

p-value = **0.268872**

3) Multiple pairwise comparison (Tukey's HSD Test)

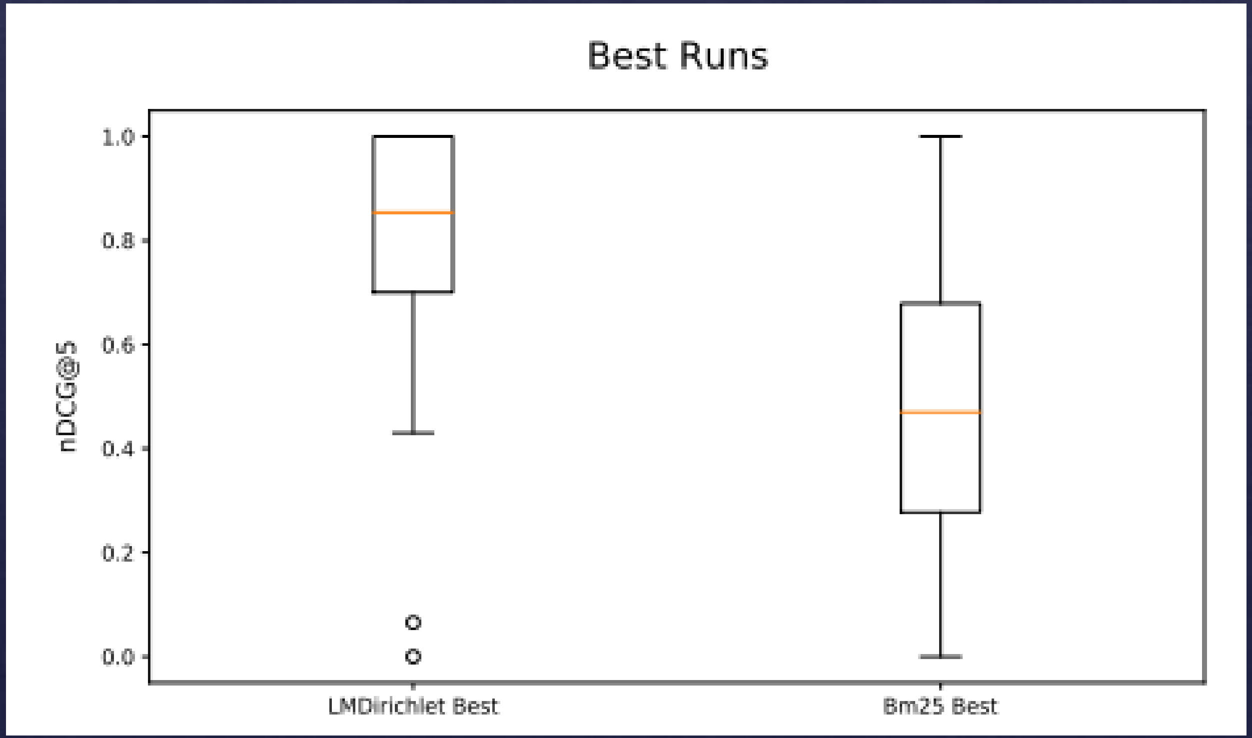
run1	run2	p-value
Dirichlet Base	Dirichlet Weights Best	0.9
Dirichlet Base	Dirichlet Synonyms Best	0.9
Dirichlet Base	Dirichlet Sentiment Best	0.81592
Dirichlet Base	Dirichlet Mix	0.372886
Dirichlet Weights Best	Dirichlet Synonyms Best	0.829768
Dirichlet Weights Best	Dirichlet Sentiment Best	0.718983
Dirichlet Weights Best	Dirichlet Mix	0.279774
Dirichlet Synonyms Best	Dirichlet Sentiment Best	0.9
Dirichlet Synonyms Best	Dirichlet Mix	0.857612
Dirichlet Synonyms Best	Dirichlet Mix	0.9

BM25 vs LMDirichlet

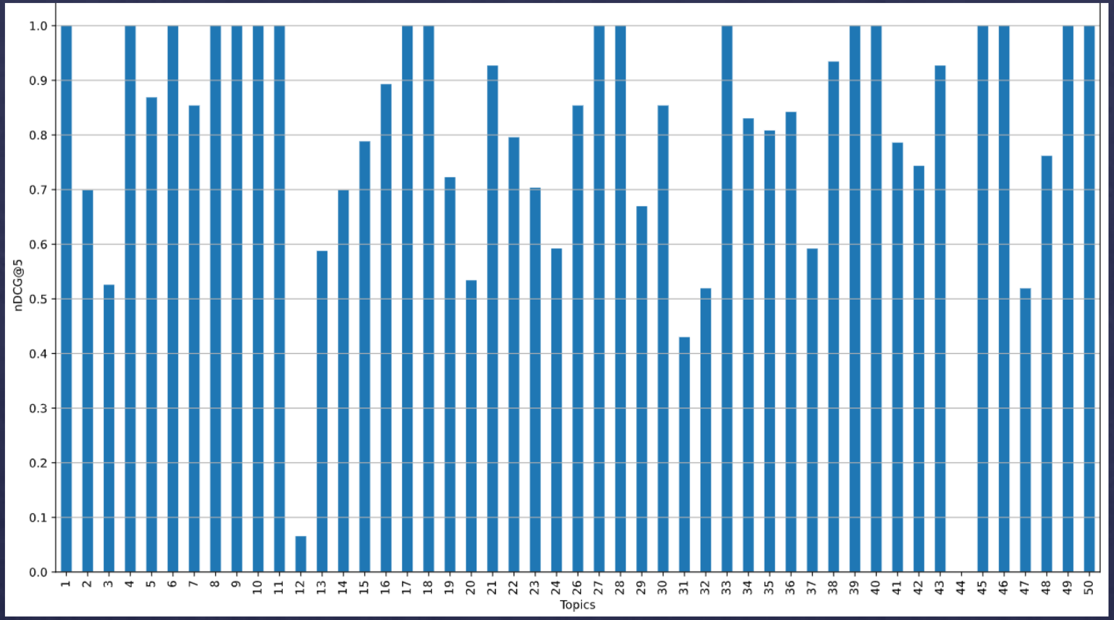
T Student Test:
p-value = **6.168497*e*−09**

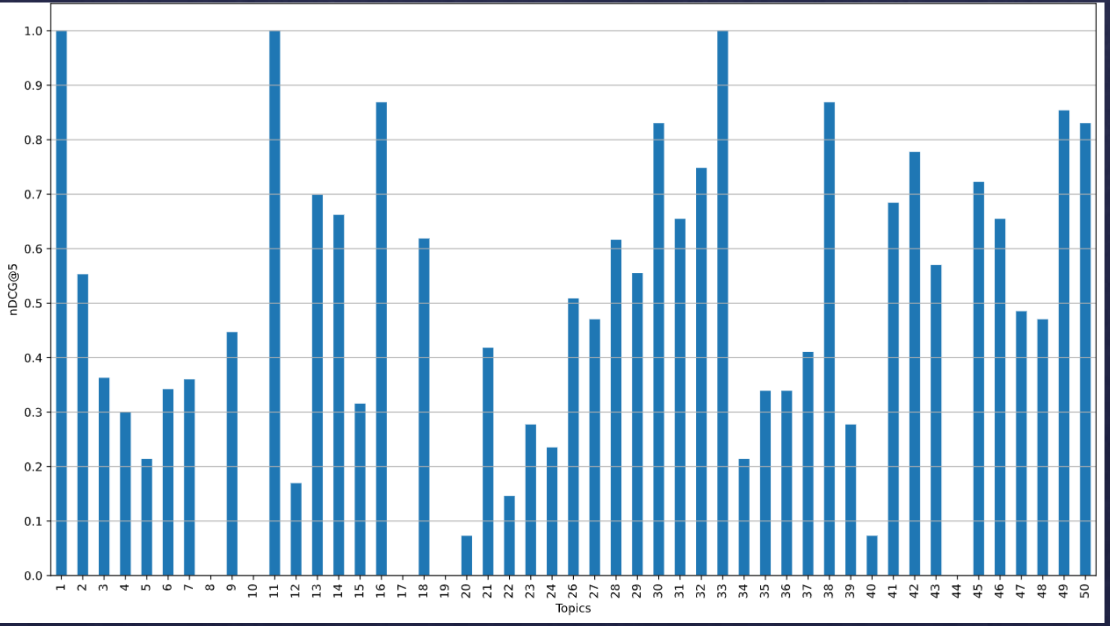
Confirm that *LMDirichlet* model is better than **BM25** for argument retrieval

Boxplots



Failure Analysis





Failure Analysis

nDCG@5 score for each topic:

- Best run using LMDirichlet
- Best run using BM25

Worst topics:

- Topic 44: Should election day be a national holiday?
- Topic 12: Should birth control pills be available over the counter?

Conclusions and future works

Conclusions:

Right weight to each field of the document

LMDirichlet better than BM25

Improvements:

Weight for each synonym of a specific word

Change formula to re-rank documents

Future works:

Machine Learning

Thanks for your attention