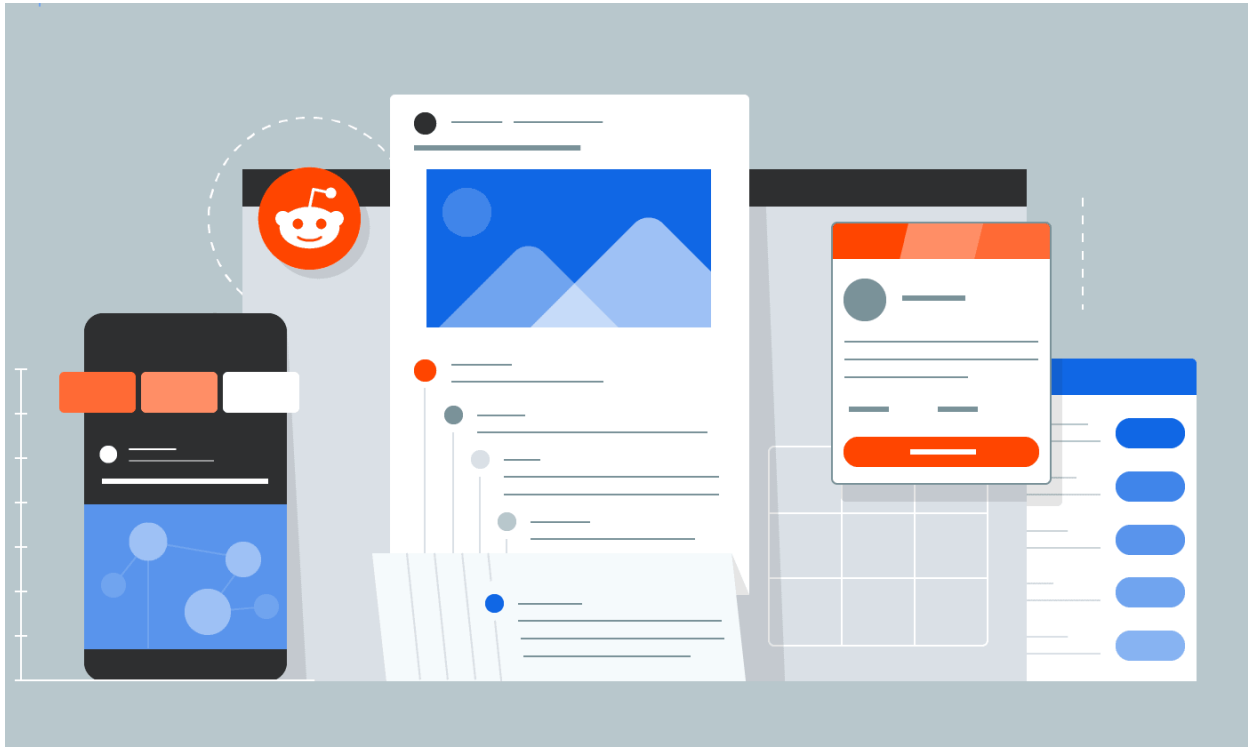
University of Milano-Bicocca
Master's Degree in Data Science
Text Mining and Search
Academic Year 2022-2023

EXTRACTIVE TEXT SUMMARIZATION AND TOPIC MODELING OVER REDDIT POSTS

Authors:
Giorgio CARBONE matricola n. 811974
Marco SCATASSI matricola n. 883823
Gianluca SCURI matricola n. 886725

reddit : the front page of the internet

- Social news aggregation & discussion website
- 100k+ communities & 430M+ posts in 2022
- **Redditors** post, comment and rate content
- **Subreddit**: domain-specific community
- **TL;DR** = "Too Long; Didn't Read" → summary for lengthy posts
- The fraction of posts containing TL;DR is decreasing



Project objectives

1. Perform **Extractive Text Summarization** on reddit posts

    - to obtain very short summaries resembling a TL;DR

2. Perform **Topic Modelling** on reddit posts

    - using Latent Dirichlet Allocation (**LDA**) and Latent Semantic Analysis (**LSA**) unsupervised algorithms

POST:
We are students and we go to college together, we have three lessons a week together. At school he normally sits at the front and I sit at the back, but recently the person I sit next to has been struggling with mental health and hasn't been in, so I moved and sit next to her most classes. For a while now we 've texted each other a few times, but outside of that, we don't really hang out at all. I see a lot of theatre, and about a week ago she said she wanted to come see a show with me. When we find our seats, mine has a pole in the way so I can't see a section of the stage unless I lean away from her. About half an hour in, this girl leans on my shoulder and starts hugging my arm, while still leaning on my shoulder. She was kind of cuddling all day, we went to an arcade earlier as well. She doesn't seem like the cuddling type of friend, and I'm very worried she has a crush on me. I don't want to ruin our friendship, I don't like her back. Should I just ignore it until she asks me? What if she thinks that was a date?

TL;DR:
I took my friend to see a show, she leant on my shoulder the whole time. I 'm not into her but I think she has a crush on me?

TOPICS:
- Topic 7: school, class, college, student, ...
- Topic 1: relationship, friend, girl, dating, ...
- ...

DATASET &
DATA EXPLORATION

Dataset: TLDRHQ

- Released in **2021**
- Posts published in 2005-2021 period
- **1,671,099 reddit posts** and their TL;DR
  - Training set → 1,590,132 instances
  - Validation set → 40486 instances
  - Test set → 40481 instances
- Attributes
  - **id** → ID of the post
  - **document** → Text of the user's post
  - **summary** → Text of the user-written TL;DR
  - **ext_labels** → Extractive labels of the post's sentences
  - **rg_labels** → Rouge scores of the post's sentences

id	train-TLDR_RS_2012-02-4890.json
document	i 'm looking for a new pair of headphones that i will carry around with me when i travel .</s><s> i do n't want to spend more than \$ 50 .</s><s> i do n't like earbuds because they do n't stay in very well .</s><s> i wear glasses so the headphones ca n't be too tight . </s><s> i 'm not an audiophile but i do appreciate quality .</s><s> i prefer over-ear style . </s><s> i 've tried [skullcandy] (https://www.amazon.com/stores/page/E0223B) , sony , and some other weird brand a while a back and so far the sony 's have the least amount of pressure but also the least amount of volume . </s><s> i ca n't turn them up because they do n't cover the ear and i 'm not that guy who walks around and forces people to listen to distorted music from headphones .
summary	want new headphones - prefer over-ear . i wear glasses so ca n't be too tight . around $ 50 . thanks !
ext_labels	[0, 0, 0, 1, 0, 1, 0, 0]
rg_labels	[0.10165, 0.11729, 0.07898, 0.36880, 0.03765, 0.15032, 0.04066, 0.10461]

Data Exploration

- Most of the posts published after 2013
- 53.8.% submissions / 46.2.% comments
- **No missing values**
- **document** and **summary** → 38K and 67K **duplicates**
  - announcements, bots messages, spam
- $\boldsymbol{compression\ rate} = \frac{document\ words\ count\ (avg)}{summary\ words\ count\ (avg)} = \mathbf{12.1}$
  - TLDRs heavily shorten the post's text

	document	summary
words count (tot)	~468M	~38M
words count (avg)	291	24
sentences count (tot)	~24M	~3.8M
sentences count (avg)	15	2
words count / sentence (avg)	20	11
unique words	~738K	~254K
compression rate		**12.1**

Data & Text Pre-processing

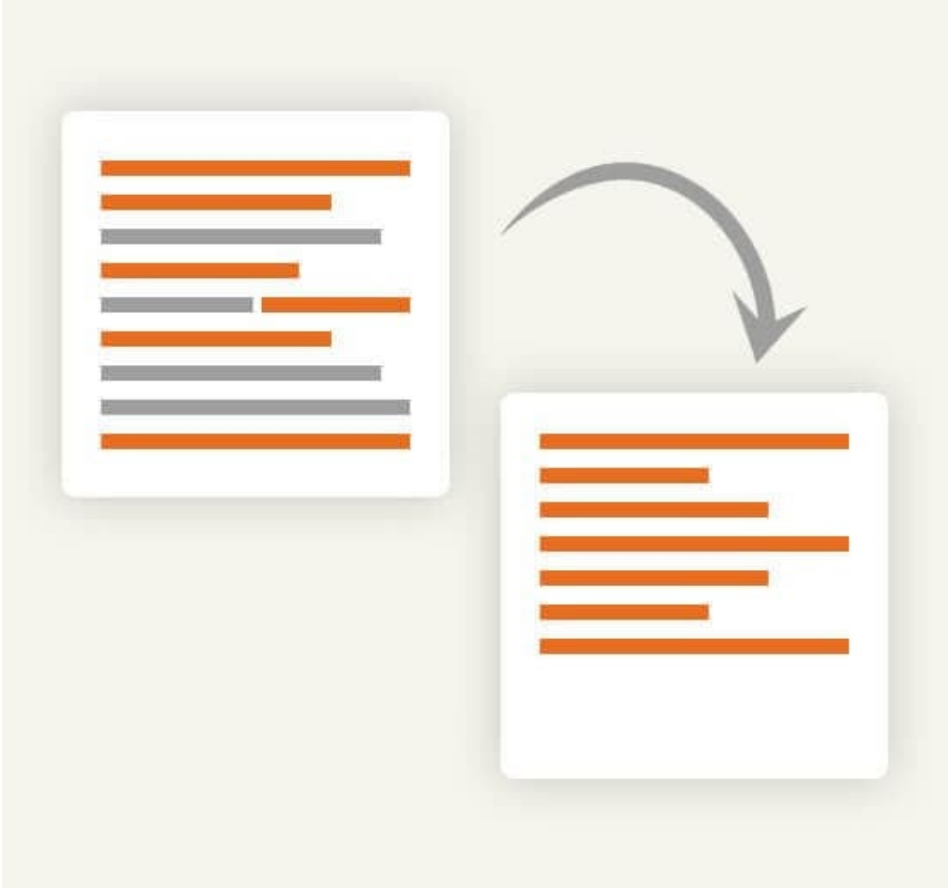
1. **Data Cleaning** → duplicates removal
2. **Sentences Splitting**
3. Text Normalization → **Text Cleaning**
4. **Text Normalization → Words and punctuation**
5. **Tokenization** → unigrams
6. **Stop-words** and 1-character words removal
7. **Lemmatization**
8. **POS Tagging**

	Text Cleaning & Tokenization	+Lemmatization & Stop-Words Removal
words count (tot)	~468M	~213M
words count (avg)	291	133
words count / sentence (avg)	20	9
unique words	~738K	~715K

TEXT SUMMARIZATION

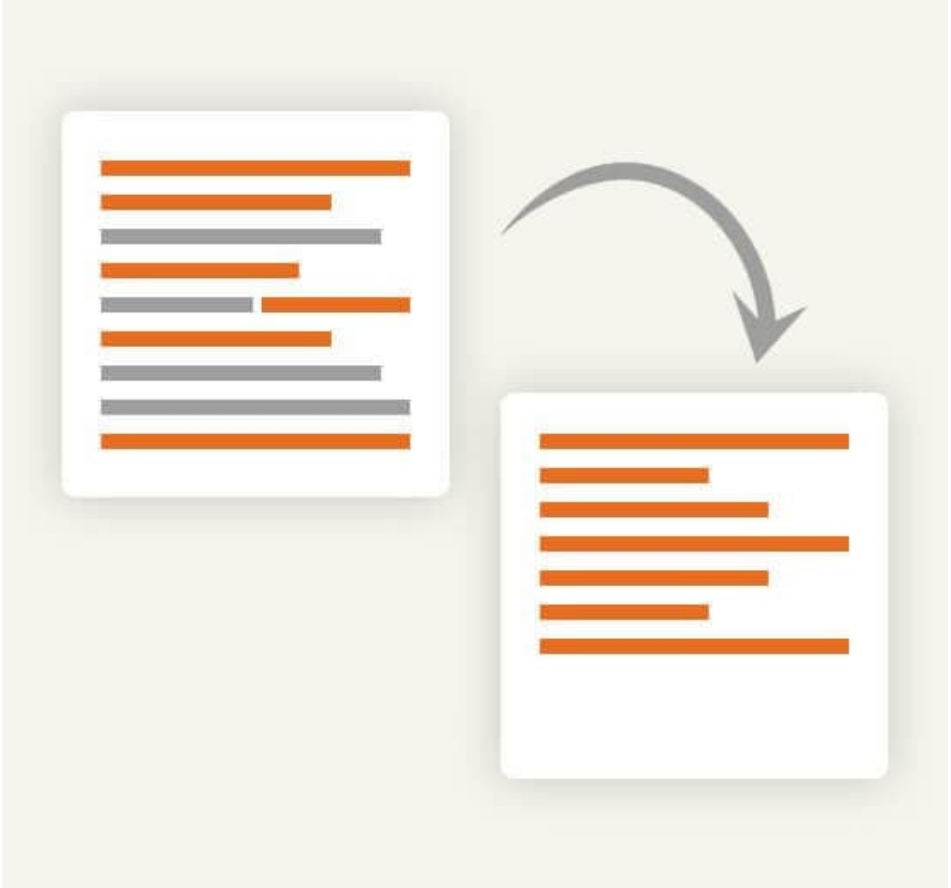
Text Summarization

- NLP task aimed at identifying and extract **the most important information** within a text
- **Many ways** to undertake this task
- Characteristics of **our approach**:
  - **single-document**
  - **generic**
  - **extractive**
- Moreover, **extreme summarization**



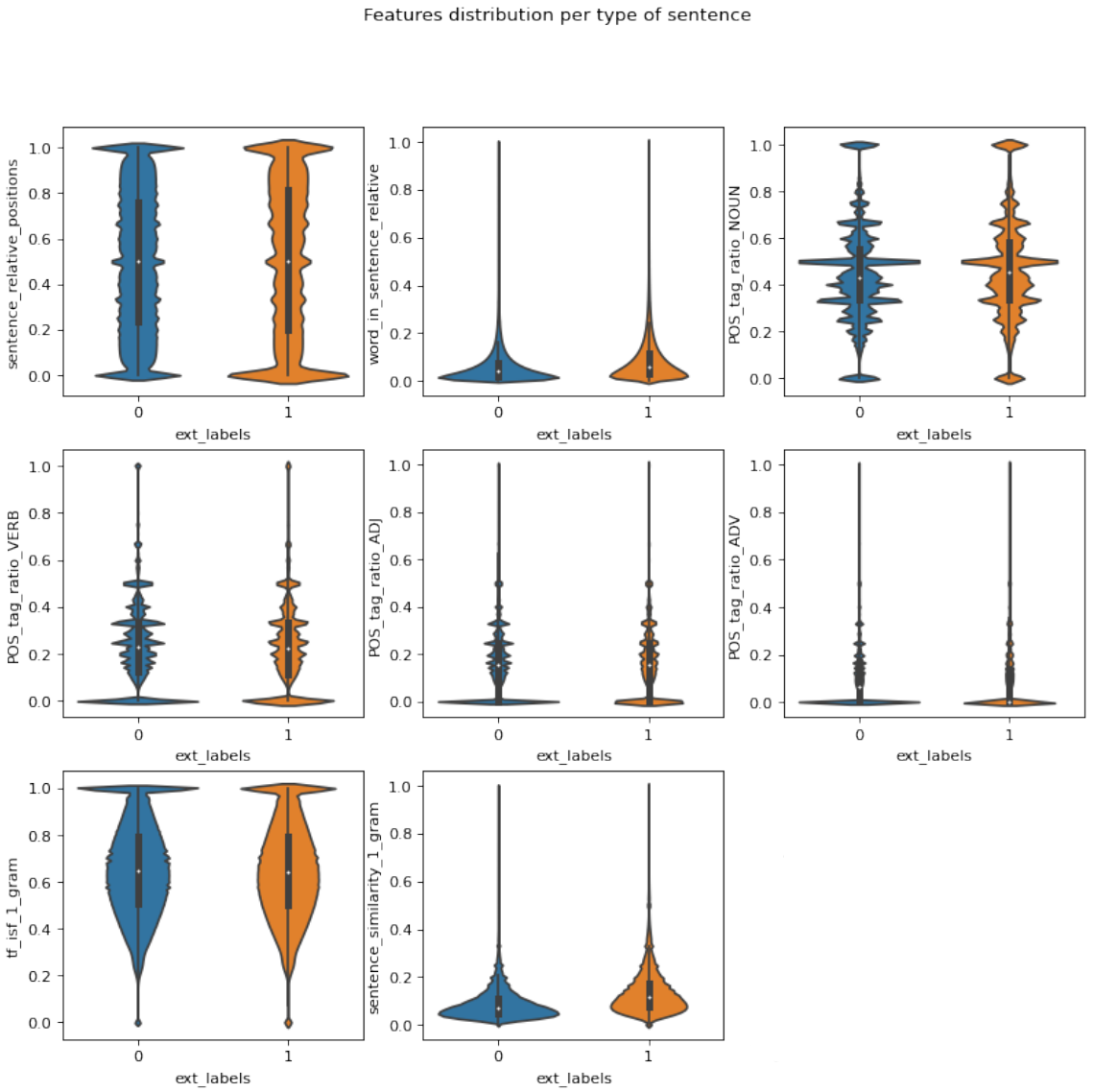
Data Pre-processing

- **Two** additional steps:
  - Removal of documents **without extractive summary**
    - 489 documents discarded
  - Removal of documents **with only one sentence**
    - 7268 documents discarded



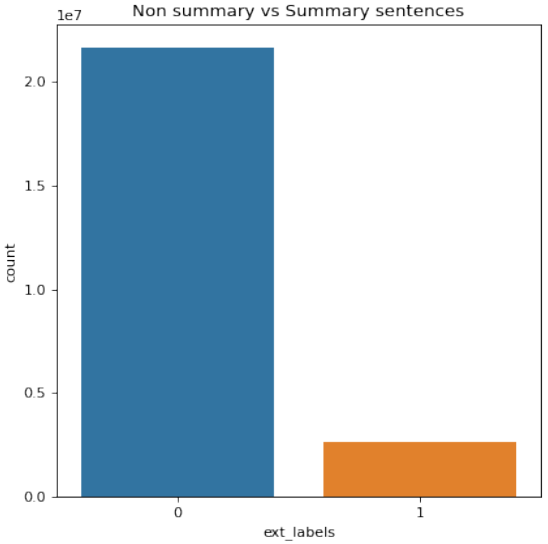
Features Matrix

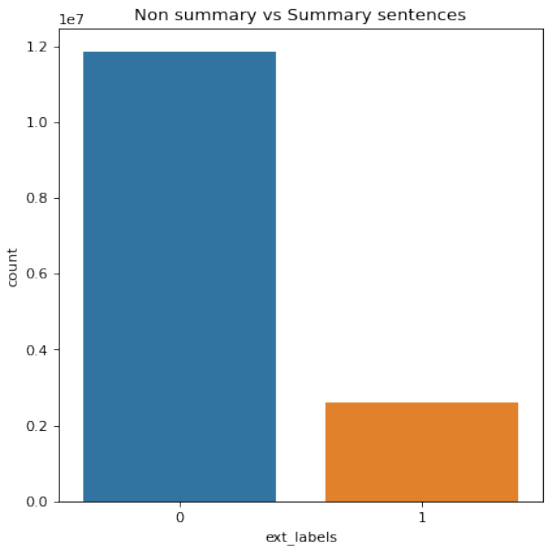
- **8 features**:
  1. Sentence **relative position**
  2. Sentence **similarity score**
  3. **Word in sentence** (relative)
  4. **NOUN** tag ratio
  5. **VERB** tag ratio
  6. **ADJ** tag ratio
  7. **ADV** tag ratio
  8. **TS-ISF**

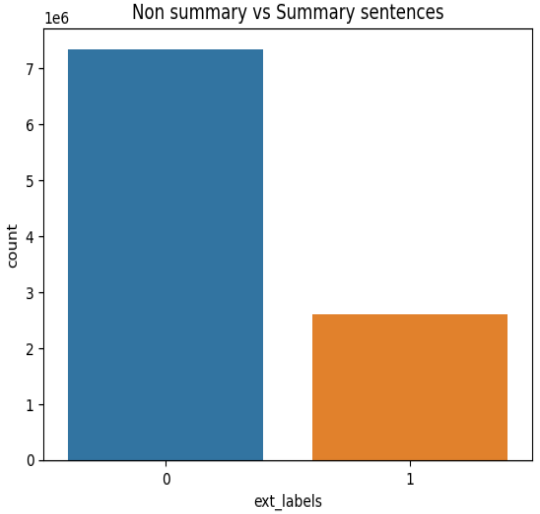


Undersampling

- **Two** methodologies:
  1. **CUR**
  2. **ENN** (Edited Nearest Neighbours)
- **Ratio** of summary to non-summary sentences:
  - **Initial**: 0.12
  - Post **CUR**: 0.22
  - Post **ENN**: 0.35







Model Selection

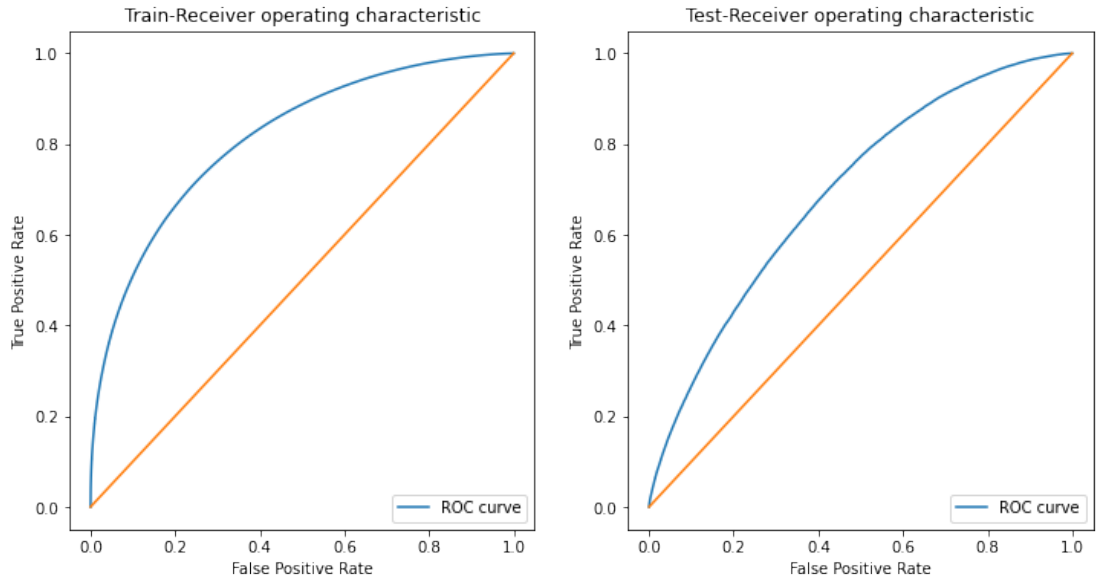
- **3-Fold** Cross Validation
- **Two** models:
  - Random Forest
  - Hist Gradient Boost
- Many parameters' configurations
  - **10 compared models**

	Mean Train F1	Mean Val F1	Mean Train Recall	Mean Val Recall	Mean Train Precision	Mean Val Precision
1	0.70	0.59	0.91	0.76	0.57	0.48
2	0.83	0.60	0.88	0.61	0.78	0.60
3	0.61	0.59	0.74	0.72	0.51	0.50
4	0.57	0.56	0.86	0.84	0.43	0.42
5	0.73	0.60	0.81	0.66	0.66	0.56
6	0.59	0.59	0.73	0.73	0.50	0.49
7	0.61	0.59	0.74	0.72	0.51	0.50
8	0.56	0.56	0.85	0.85	0.41	0.41
9	0.73	0.60	0.81	0.66	0.66	0.55
10	0.78	0.60	0.94	0.69	0.67	0.52

Model Selection

- Final Model
  - **Hist Gradient Boost**:
    - 500 max iterations
    - 255 bins
    - 0.05 as learning rate
    - $\{0 : 1, 1 : 3\}$ as class weights

	Mean Train F1	Test F1	Mean Train Recall	Test Recall	Mean Train Precision	Test Precision
6	0.59	0.24	0.73	0.72	0.50	0.16



Summary Selection

- Three length summaries
  - One - sentence
    - extreme summarization
  - Two - sentences
  - Sqrt – sentences

$$MMR = \arg \max_{D_i \in R \setminus S} \left[\lambda score(D_i) - (1 - \lambda) \max_{D_j \in S} \mathrm{Sim}(D_j, D_i) \right]$$

- D_i, sentences in a document D
- R, set of sentences not yet chosen
- S, current summary
- $\lambda = 0.05$

Summary Evaluation

- **Comparable** performance
  - Despite **lower** complexity
- Explainability

- More sentences
  - **Slightly increase** in RG scores

Model	RG-1(%)	RG-2(%)	RG-L(%)
BertSumExt	28.40	11.35	21.38
One-sentence Hist Gradient Boost	21.99	7.06	17.15
Two-sentences Hist Gradient Boost	24.34	7.83	17.44
Sqrt-sentences Hist Gradient Boost	25	8	18

Example 1

- Example in which the model **couldn't capture** the crucial information in all summaries

Original Post	[light] : casual confessions i met this german girl through my best friend , we drank at my house and then head to the club . we party , we dance , we have a good time . i then proceed to go to the couch and sit down , i was tired . my best friend and the other girl we were with go to the bar to order something . the german girl then comes and sits besides me , we talk and then i put my arm around her and she comes closer . i could n't believe it ! this is one of the most beautiful girls i 've ever met . she 's tall , blonde , blue eyes ... perfect . i start to caress her arm , and then it happened . she goes in for the kiss . i ca n't believe what is just happening ! we make out for a solid 10 - 15 . but it goes beyond that . when i had my arm around her , i felt comfortable , and when she leaned in for the kiss , i felt like the whole world stopped , i felt happy ... i forgot about all my problems ... like if she was the one . i remember when i got home i was drunk and happy , but then it sank in that i 'm never going to see this girl again . then i got really sad about it , like if all of this was just a happy coincidence that was n't really meant to happen . i woke up today , feeling kinda hangover . but all that i could think about was her . her accent , her face , those beautiful blue eyes and they way she smiled at me . all i can think off is her , and the fact that i 'm never going to see her again it 's really hurting me . i go back and remember when we were kissing . it was romantic , how she hold me . the way she would stop and look at me , then continue to kiss me . is like we were in love , an impossible love .
TL;DR	the love of my life may very well be on a plane back to germany , and i 'm never going to see her again , and that fact is destroying me .
Extractive reference summary	all i can think off is her , and the fact that i 'm never going to see her again it 's really hurting me .
One sentence summary	i go back and remember when we were kissing .
Two sentence summary	this is one of the most beautiful girls i 've ever met . i go back and remember when we were kissing ..
Sqrt sentence summary	this is one of the most beautiful girls i 've ever met . but all that i could think about was her . i go back and remember when we were kissing . it was romantic , how she hold me . the way she would stop and look at me , then continue to kiss me .

Example 2

- The **one sentence** summary is **not representative** of the TL;DR.
- The **two sentence one captures the most relevant information** inside the post

Original Post	so i live at home with my parents and we all work full time . my parents both work day shift while my schedule varies . over the course of the past year i 've noticed that i 'm honestly quiet the jerk and i 'd like to change that . i 'm closest to my parents because i do n't have much of an extended family that 's worth writing about . so sometimes my parents do things that bug me . my mom for examples tends to burp a lot when she eats and it really disgusts me . i 'm not sure how i can go about saying " hey , that 's really disgusting can you not " without saying it like that . my dad tends to eat with his mouth open a lot which again , is rather annoying at a restaurant . the same applies here , how do i say something without being a jerk . one thing they both do is smoke and i want them to quit because of obvious reasons . i know i ca n't force them to stop but it hurts me knowing that everytime they have a smoke it takes a couple minutes away from them . i love my parents and these are just a few of things that seem to bug me . i guess i just want some advice on how to be a better person to my parents .
TL;DR	been a jerk to my parents , how do i fix it .
Extractive reference summary	the same applies here , how do i say something without being a jerk .', ' i guess i just want some advice on how to be a better person to my parents .
One sentence summary	so sometimes my parents do things that bug me .
Two sentence summary	so sometimes my parents do things that bug me . the same applies here , how do i say something without being a jerk .
Sqrt sentence summary	so sometimes my parents do things that bug me . my mom for examples tends to burp a lot when she eats and it really disgusts me . the same applies here , how do i say something without being a jerk .

Example 3

- A **good** one sentence summary.
- Two and Sqrt summaries **add non relevant information**
- The extractive reference summary is **wrong**.
  - Examples like this one, negatively affected the performance of our model

Original Post	perk sends this after over a week of back and forth . " thank you for your response . we have now closed any additional accounts in your household and have unflagged the accounts linked xxxxxxxxx happy perking ! please feel free to contact us if you have any additional questions or concerns and we will be happy to assist you . thank you for choosing perk ! i redeemed for my account yesterday and i got this your reward has been cancelled your perk account has been found in violation of our terms of service by our fraud detection software . your account has been flagged and your rewards have been cancelled . why bother unflagging me if you are just going to say " just kidding they are all cancelled " perk is really pissing me off . i refuse to use it until they figure their crap out .
TL;DR	unflagged my account and then cancelled all rewards i tried to redeem for .
Extractive reference summary	i redeemed for my account yesterday and i got this
One sentence summary	your account has been flagged and your rewards have been cancelled .
Two sentence summary	your account has been flagged and your rewards have been cancelled . perk is really pissing me off .
Sqrt sentence summary	thank you for your response . your account has been flagged and your rewards have been cancelled . perk is really pissing me off .

Example 4

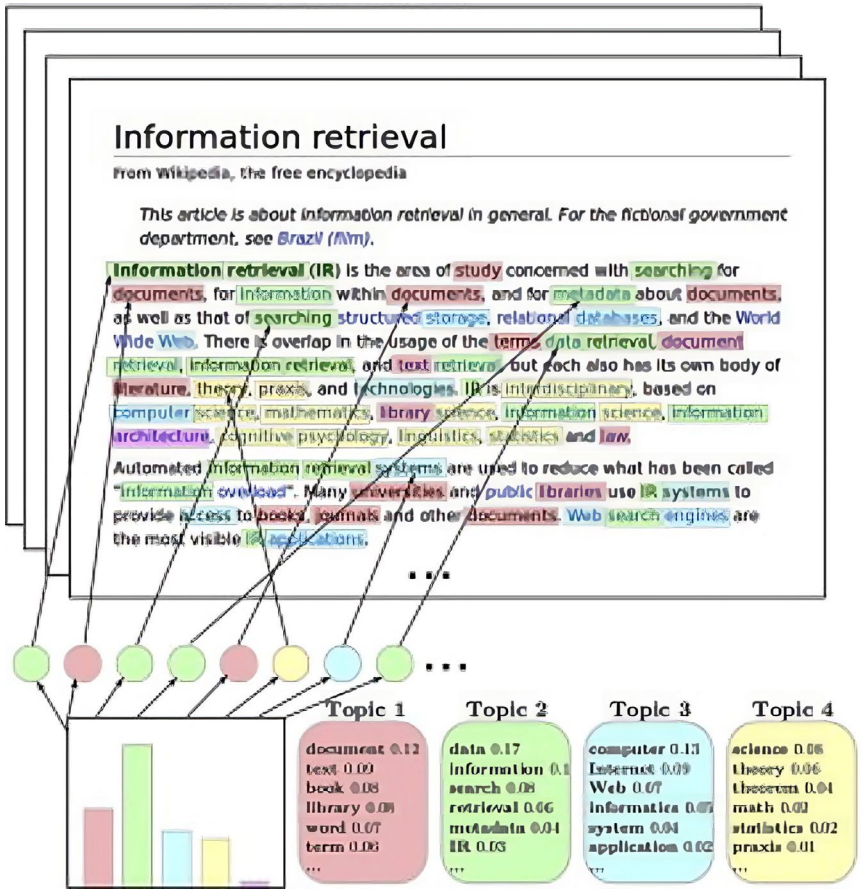
- The one sentence summary **captures quite well all the information** inside the TL:DR
  - it is equal to the reference extractive summary
- Two and Sqrt summaries **add non relevant information**

Original Post	even with deniability , the problem is that something has hooked into the account to store these scripts in their cloud system tied to the account and pull it down whenever a sync is ran . why the system is able to create these files out in a standard malware target location , instead of keeping contained to user appdata folders for config files of macros and lighting profiles is beyond me . sure , you could craft a macro to craft those files if you really felt like it , but that would be something that requires a button press , not an automatic thing on sync .
TL;DR	someone has figured out how to hook some scripts into the cloud storage system tied to razer 's accounts and it will pull down automatically with no way for you to remove it .
Extractive reference summary	even with deniability , the problem is that something has hooked into the account to store these scripts in their cloud system tied to the account and pull it down whenever a sync is ran .
One sentence summary	even with deniability , the problem is that something has hooked into the account to store these scripts in their cloud system tied to the account and pull it down whenever a sync is ran .
Two sentence summary	even with deniability , the problem is that something has hooked into the account to store these scripts in their cloud system tied to the account and pull it down whenever a sync is ran . why the system is able to create these files out in a standard malware target location , instead of keeping contained to user appdata folders for config files of macros and lighting profiles is beyond me .
Sqrt sentence summary	even with deniability , the problem is that something has hooked into the account to store these scripts in their cloud system tied to the account and pull it down whenever a sync is ran .

TOPIC MODELING

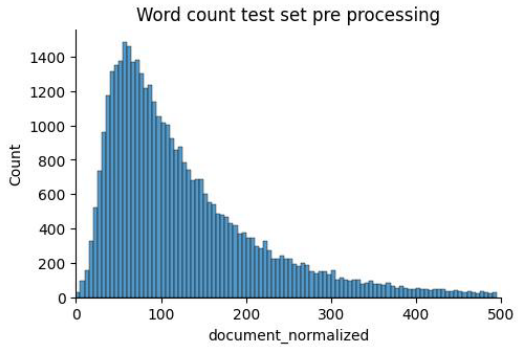
Topic Modeling

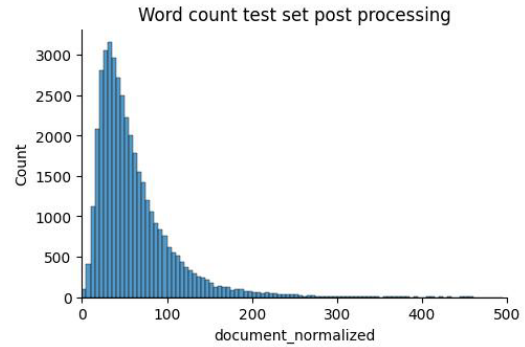
- NLP task aimed at detecting word and phrase patterns within them
- Groups words into different clusters, where each word in the cluster is likely to occur “more” for the given **topic**
- The results of this task are **3 components**:
  - **list of topics**
  - probability of **words for each topic**
  - distribution of **topics for each document**
- Latent Semantic Analysis (**LSA**) and Latent Dirichlet Allocation (**LDA**)
- Apply these techniques to the test set of the TLDRHQ dataset



Exploration and pre-processing

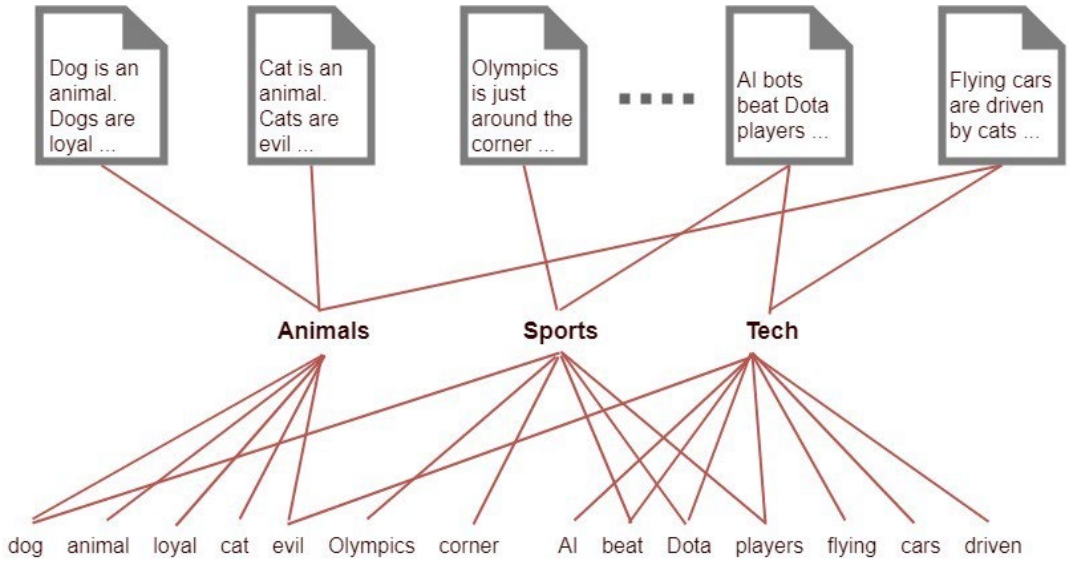
- Important that documents contain only words with **high semantic value**
- Actions performed:
  - **Verb and noun filtering** based on POS tagging (90444 -> 75879 terms)
  - **Common** (>50% of documents) and **rare** (<5 documents) **words removal (Zipf's law)** (75879 -> 75861 terms)
  - **custom stop-words** ['time', 'something', 'going', 'year', 'week', 'month', 'day', 'get', 'got', 'want', ...] (75861 -> 13536)

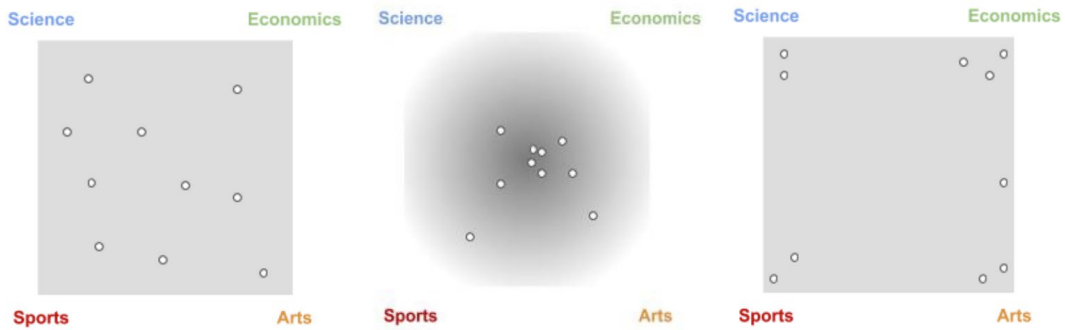




LDA: introduction

- Generative probabilistic model of a corpus
- **Two key assumptions:**
  - documents are a mixture of topics
  - topics are a mixture of words
- Both follow the **Dirichlet distribution**
- **Hyper-parameters**: α, β and k
- **Steps:**
  - BOW representation
  - LdaMulticore() and CoherenceModel() of Gensim

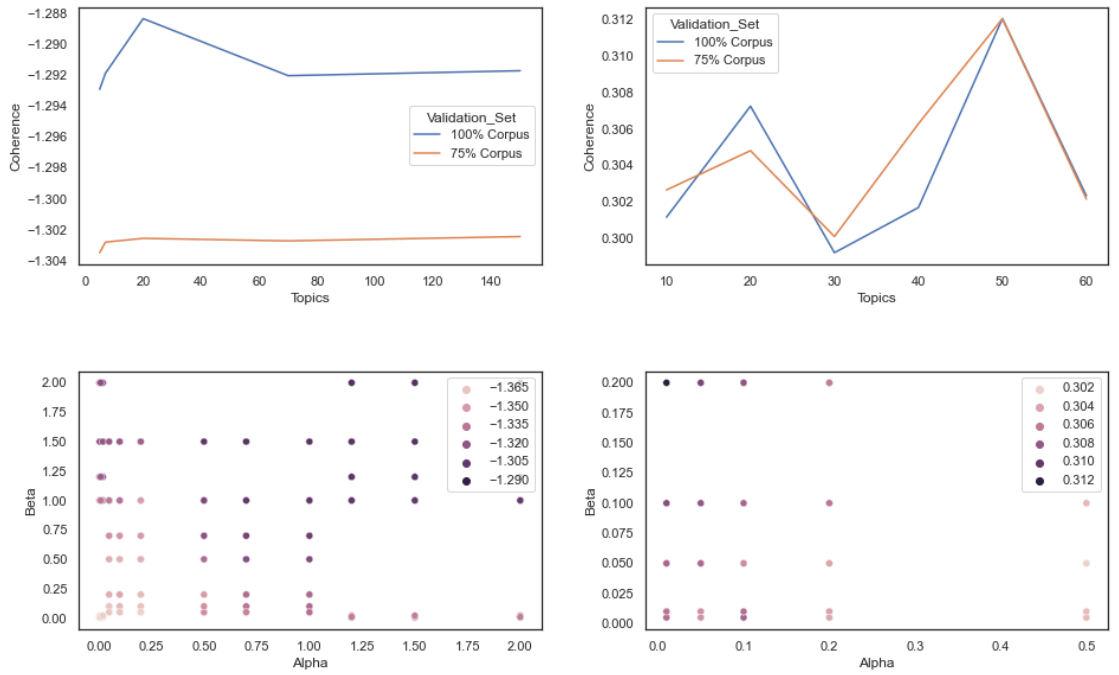




$\alpha = 1$ $\alpha > 1$ $\alpha < 1$

LDA: hyper-parameters tuning

- Search grid maximizing coherence (intrinsic evaluation metric)
- **UMass**
  - $\alpha > 1$
  - $\beta > 1$
  - $k = 20$
- **c_v**
  - $\alpha = 0.01$
  - $\beta = 0.2$
  - $k = 20$ or $k = 50$



LDA: results

- Topics obtained are **not well distinguished** and many of them have **overlapped** terms
- Even with other hyper-parameters sets (smaller β) the result doesn't improve probably due to the little number of words for every document
- Possible solutions:
  - Expand ranges grid search (more topics)
  - Learn topics on the entire dataset
  - Different coherence measures
  - Further pre-processing operations

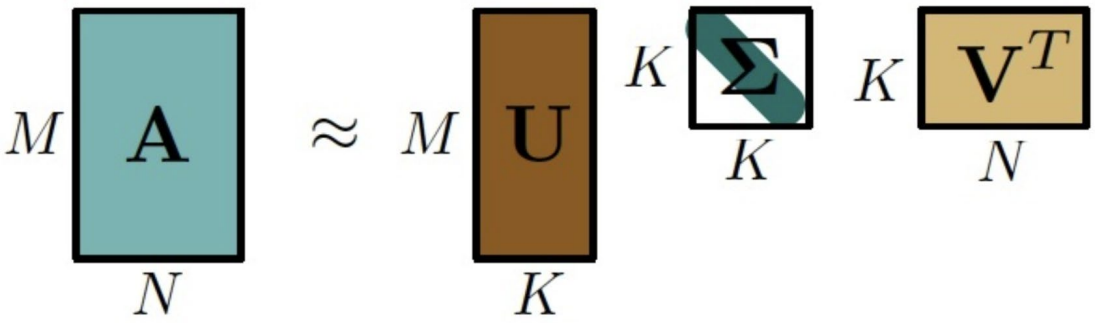
Doc	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8
0	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008	0.0008
1	0.4687	0.0000	0.0993	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.0005	0.0005	0.0005	0.0005	0.7440	0.0005	0.0005	0.2300	0.0005
3	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
4	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004
5	0.0001	0.0001	0.0001	0.0001	0.0001	0.1694	0.0001	0.0001	0.0001
6	0.0001	0.0001	0.0001	0.0001	0.0001	0.9930	0.0001	0.0001	0.0001
7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
8	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
9	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004

Topic	Top10Word	Score
0	people	0.0065
0	make	0.0065
0	think	0.0059
0	way	0.0051
0	work	0.0049
0	need	0.0045
0	take	0.0044
0	see	0.0043
0	Friend	0.0042
0	feel	0.0041

Top10Words	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8
0	people	make	think	work	think	feel	way	make	make
1	make	think	feel	way	lot	way	think	work	see
2	think	feel	people	need	way	think	make	think	think
3	way	lot	make	people	make	let	people	people	way
4	work	friend	anything	think	people	make	see	point	lot
5	need	point	need	take	feel	anything	friend	friend	friend
6	take	work	point	point	see	lot	everything	way	made
7	see	way	see	make	friend	friend	getting	life	people
8	friend	life	way	getting	anything	see	lot	anything	take
9	feel	see	take	help	need	people	anything	help	let

LSA: introduction

- Assumes that similar documents will contain approximately the same distribution of word frequencies for certain words.
- Consist in a **truncated Singular Value Decomposition** of the document-term matrix in Document-Topic and Term-Topic ($A = U\Sigma V^T$)
- Requires the number of components/topic ($K = 20$)



- **Steps:**
  - Tf-idf representation
  - randomized_svd() of Sklearn

LSA: results

- These results are much more **meaningful** than the LDA analysis
- Topics are now **well defined and distinguished**

Doc	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8
0	0.0003	-0.0008	0.0001	0.0001	-0.0011	0.0002	0.0006	0.0004	0.0000
1	0.0104	0.0074	0.0009	0.0006	-0.0025	0.0045	-0.0075	-0.0049	-0.0048
2	0.0031	-0.0006	-0.0025	-0.0048	0.0009	0.0047	0.0034	0.0121	-0.0060
3	0.0062	-0.0034	-0.0095	-0.0028	-0.0081	0.0021	0.0006	0.0022	0.0059
4	0.0031	-0.0003	0.0012	-0.0044	-0.0003	0.0030	-0.0055	-0.0012	0.0003
5	0.0064	-0.0032	0.0026	-0.0011	0.0006	0.0026	-0.0027	0.0021	0.0018
6	0.0041	-0.0073	-0.0001	0.0087	0.0014	0.0036	0.0028	-0.0044	0.0009
7	0.0042	-0.0039	0.0073	-0.0054	0.0091	-0.0140	0.0054	-0.0137	-0.0015
8	0.0034	0.0002	0.0019	-0.0041	0.0038	0.0085	-0.0018	0.0001	0.0051
9	0.0017	-0.0010	0.0016	-0.0003	-0.0052	-0.0012	-0.0003	-0.0006	-0.0041

Topic	Top10Word	Score
0	feel	0.12
0	friend	0.12
0	make	0.12
0	think	0.12
0	people	0.11
0	see	0.11
0	way	0.11
0	lot	0.1
0	relationship	0.1
0	work	0.1

Top10Words	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4	Topic_5	Topic_6	Topic_7	Topic_8
0	feel	relationship	went	game	game	game	school	school	guy
1	friend	friend	home	play	hundred	family	people	class	girl
2	make	girl	house	playing	money	parent	class	college	people
3	think	dating	car	player	play	life	friend	student	hundred
4	people	told	room	played	work	play	college	experience	money
5	see	boyfriend	came	friend	buy	playing	game	started	car
6	way	feeling	minute	guy	pay	mom	student	looking	make
7	lot	met	took	team	hour	kid	everyone	girl	pay
8	relationship	talk	door	girl	playing	love	girl	grade	buy
9	work	asked	hour	fun	friend	dad	kid	help	date

LSA: example

- Topic 7: **school**
- Topic 3: **games**
- Topic 13: **advices**

Top10Words	Topic_3	Topic_7	Topic_13
0	game	school	post
1	play	class	advice
2	playing	college	question
3	player	student	help
4	played	experience	anyone
5	friend	started	twenty
6	guy	looking	hundred
7	team	girl	looking
8	girl	grade	team
9	fun	help	thirty

Document	Topic
i can not for the life of me find a school for wing chun , and i am very eager to learn . i know that bad habits can come from learning online but i am getting restless . so if any one would like to help a (hopefully) soon to be chunner out , find a school near pleasanton , ca . lineage is n't a big concern of mine right now	Topic_7
i have noticed lately that while i may not be interested in a game for its gameplay i want to dig into the story and lore as much as possible . as an example i cant stand the puzzle platforming or multiplayer aspects of splatoon but the lore and characters are super interesting to me ... i want to enjoy those parts of it but have no desire to actually play the game .	Topic_3
since im using a ipad pro with a apple pencil for my work and my studies i had this cool idea of sketching little things and put them later on as wallpaper on my desktop but i ve tested some free apps like adobe sketch but sadly in all these apps , the screen resolution / format did nt fit my dektops one . so my sketch gets either cut off or i have like two big black bars on my desktop . so maybe you know a sketching / drawing app where i can change the resolution or where the format fits my desktop . im using a 1080p monitor btw . i appreciate every suggestion :d	Topic_13

/ Bibliography

1. Sajad Sotudeh et al. “TLDR9+: A Large Scale Resource for Extreme Summarization of Social Media Posts”. In: (Nov. 2021), pp. 142–151. DOI: 10.18653/v1/2021.newsum-1.15. URL: https://aclanthology.org/2021. newsum-1.15
2. Reddit - Dive into anything. URL: https://www.reddit.com.
3. Sahil Patel. Reddit Claims 52 Million Daily Users, Revealing a Key Figure for Social-Media Platforms. Ed. by The Wall Street Journal. URL: https://www.wsj.com/articles/reddit-claims-52-million-daily-users-revealing-a-key-figure-for-social-media-platforms-11606822200. (posted: Dec. 1, 2020).
4. Reddit Staff Announcements. Revealing This Year’s (2022) Reddit Recap. Ed. by upvoted: The Official Reddit blog. URL: https://www.redditinc.com/blog/reddit-recap-2022-global. (posted: Dec. 8, 2022)
5. ir@Georgetown - Home, ed. The Georgetown University Information Retrieval Lab. URL: https : / / ir . cs . georgetown.edu/. (accessed: January 16, 2023)
6. Sajastu. sajastu/reddit collector: Reddit Collector and Text Processor. URL: https://github.com/sajastu/reddit_collector.
7. Vinicius Camargo da Silva, Jo ao Paulo Papa, and Kelton Augusto Pontara da Costa. Extractive Text Summarization Using Generalized Additive Models with Interactions for Sentence Selection. 2022. arXiv: 2212.10707 [cs.CL].

/ Bibliography

8. Kam-Fai Wong, Mingli Wu, and Wenjie Li. “Extractive Summarization Using Supervised and Semi-Supervised Learning”.
In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, UK: Coling 2008 Organizing Committee, Aug. 2008, pp. 985–992. URL: https://aclanthology.org/C08-1124.

9. Alexander Dlikman and Mark Last. “Using Machine Learning Methods and Linguistic Features in Single-Document Extractive Summarization”. In: DMNLP@PKDD/ECML. 2016.

10. Guillaume Lemaˆıtre, Fernando Nogueira, and Christos K. Aridas. “Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning”. In: Journal of Machine Learning Research 18.17 (2017), pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.

11. David M Blei, Andrew Y Ng, and Michael I Jordan. “Latent dirichlet allocation”. In: Journal of machine Learning research 3.Jan (2003), pp. 993–1022.

12. Scott C Deerwester et al. Computer information retrieval using latent semantic structure. US Patent 4,839,853. June 1989.