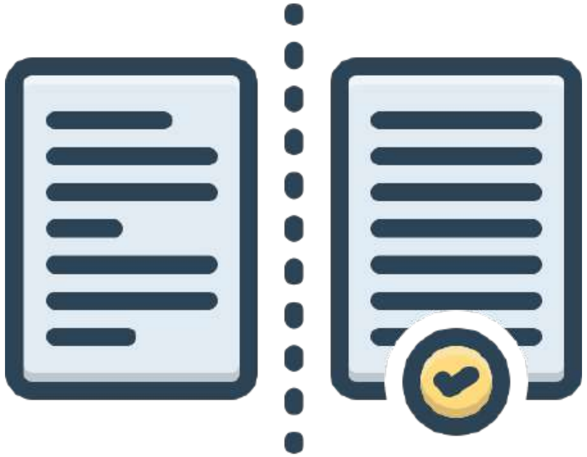
Text Data Mining
For Business Decisions

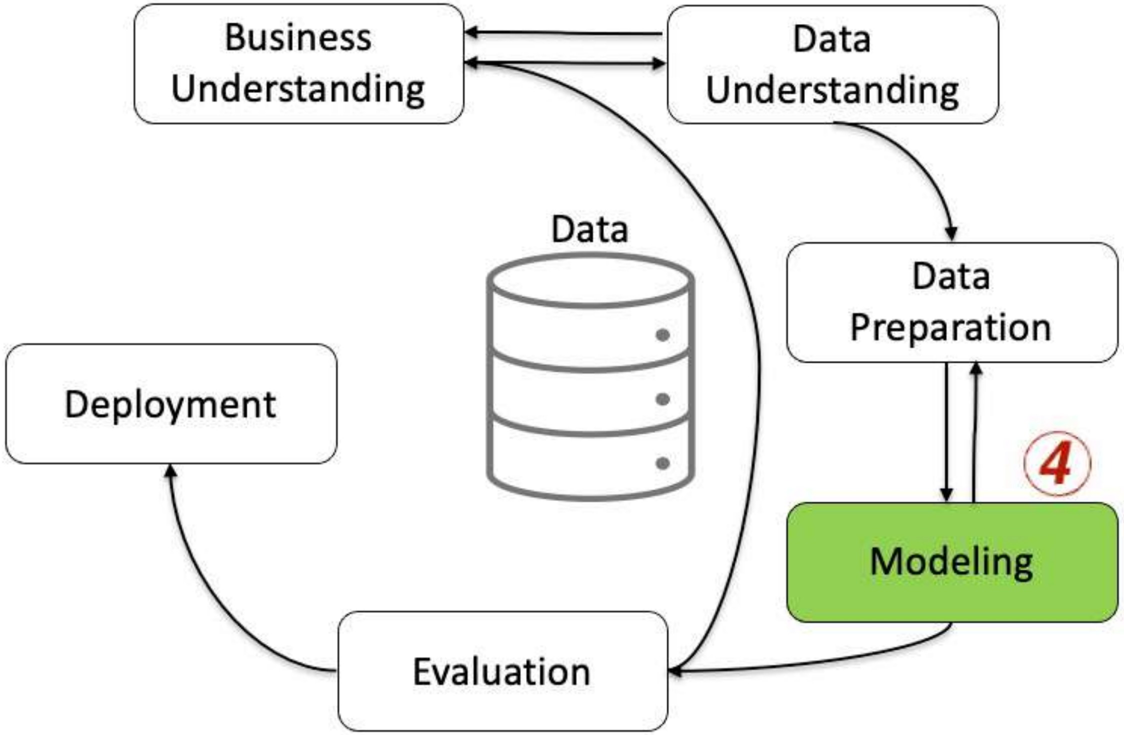
Module 9

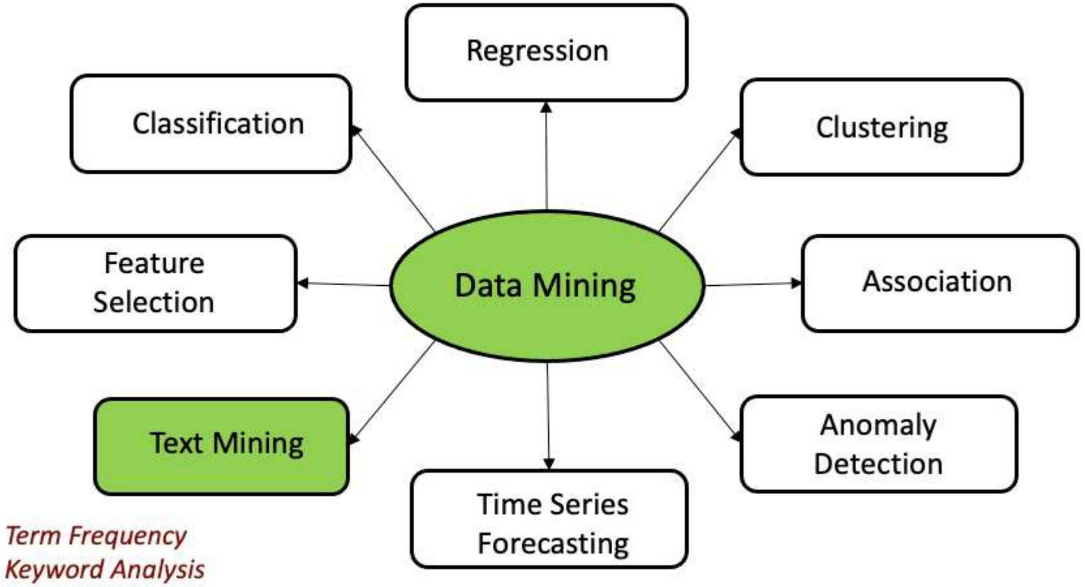
Text Similarity Scoring

How do we compare the similarities of two text documents?



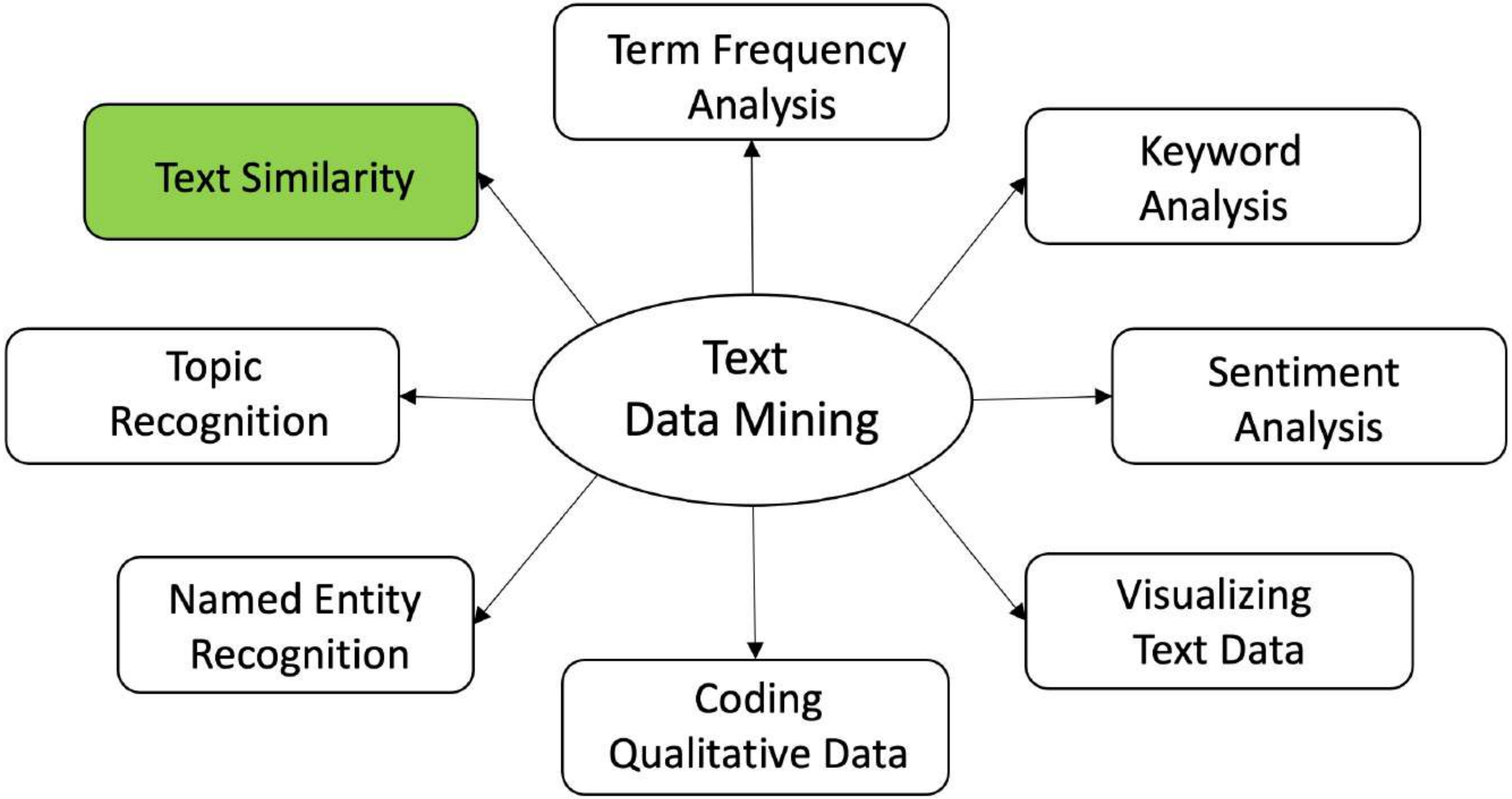
Data Mining- Continuing with Model-Making





Term Frequency
Keyword Analysis
Sentiment Analysis
Similarity Scoring

Text Similarity Scoring



What is Text Similarity Scoring?

- Take, for example, these three texts:

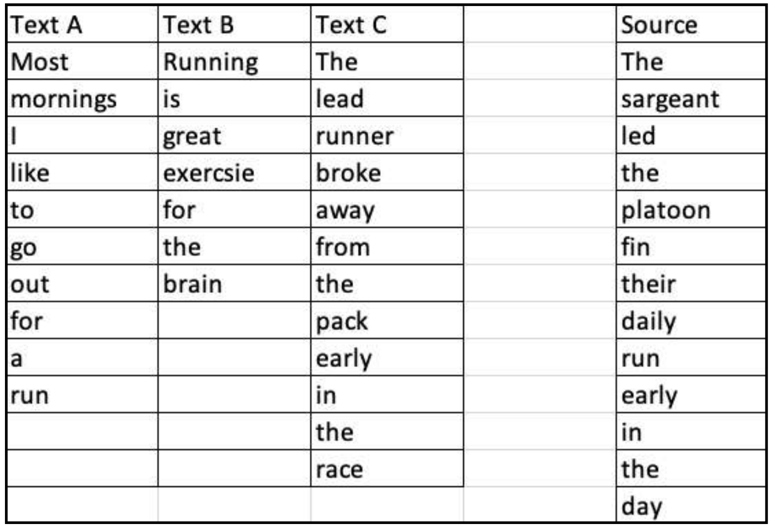
  A - Most mornings, I like to go out for a run.

  B - Running is an excellent exercise for the brain.

  C - The lead runner broke away from the pack early in the race.

- We want to compare these statements against this one-sentence document:
  - The sergeant led the platoon in their daily run early in the day.
- Which of the three texts above is most similar to the fourth text?

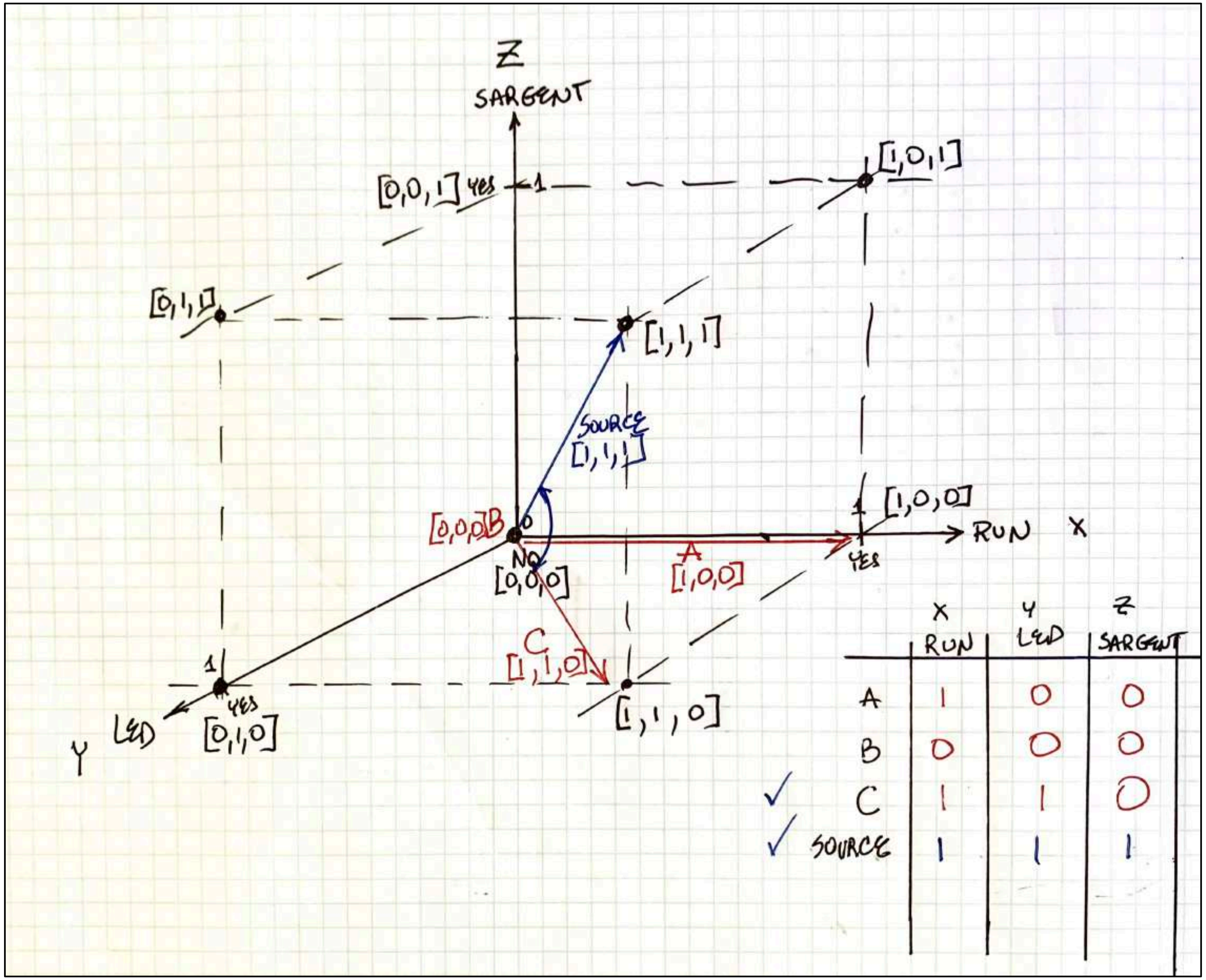
- The three sentences are the target, and the fourth is our source. In the first step, the algorithm extracts all the terms and produces a Bag-of-Words for each (as we did in early chapters).



Text A	Text B	Text C		Source
Most	Running	The		The
mornings	is	lead		sargeant
I	great	runner		led
like	exercsie	broke		the
to	for	away		platoon
go	the	from		fin
out	brain	the		their
for		pack		daily
a		early		run
run		in		early
		the		in
		race		the
				day

TF-IDF Scoring

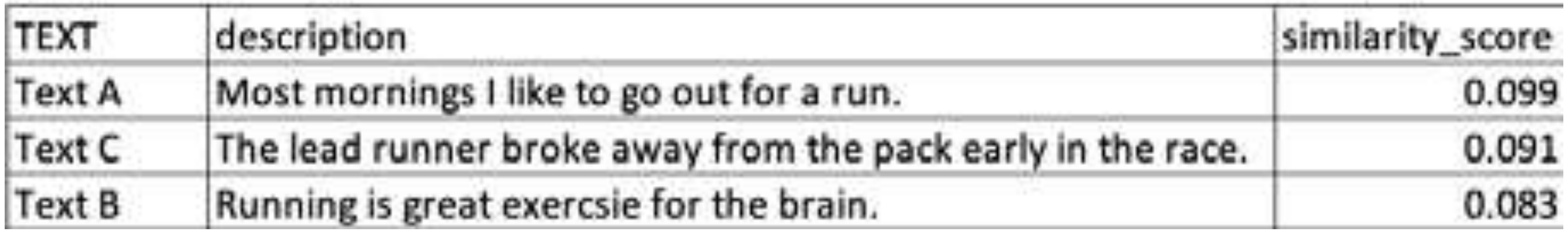
- In the next step, the algorithm removes all the stop words (*I*, *to*, *a*).
- Then tokenizes and lemmatizes all terms (*run* and *runner* get converted to *run*).
- The TF, or *term frequency*, is computed next (essentially, it performs a word frequency analysis).
  - But if some words are too frequent, they may not be too interesting (like the word "lawyer" in contracts: we all know they will be there, so they are commonplace and should be downplayed).
  - The algorithm downplays them by using the inverse of the frequency (the IDF part). We are left with lists of words and their inverse frequencies.
- Now we compare the list of words and their score to see if they have words in common and compute a common score normalized to 1 (the *cosine similarity score*).



	X RUN	Y LED	Z SARGENT
A	1	0	0
B	0	0	0
✓ C	1	1	0
✓ SOURCE	1	1	1

TF-IDF Scoring

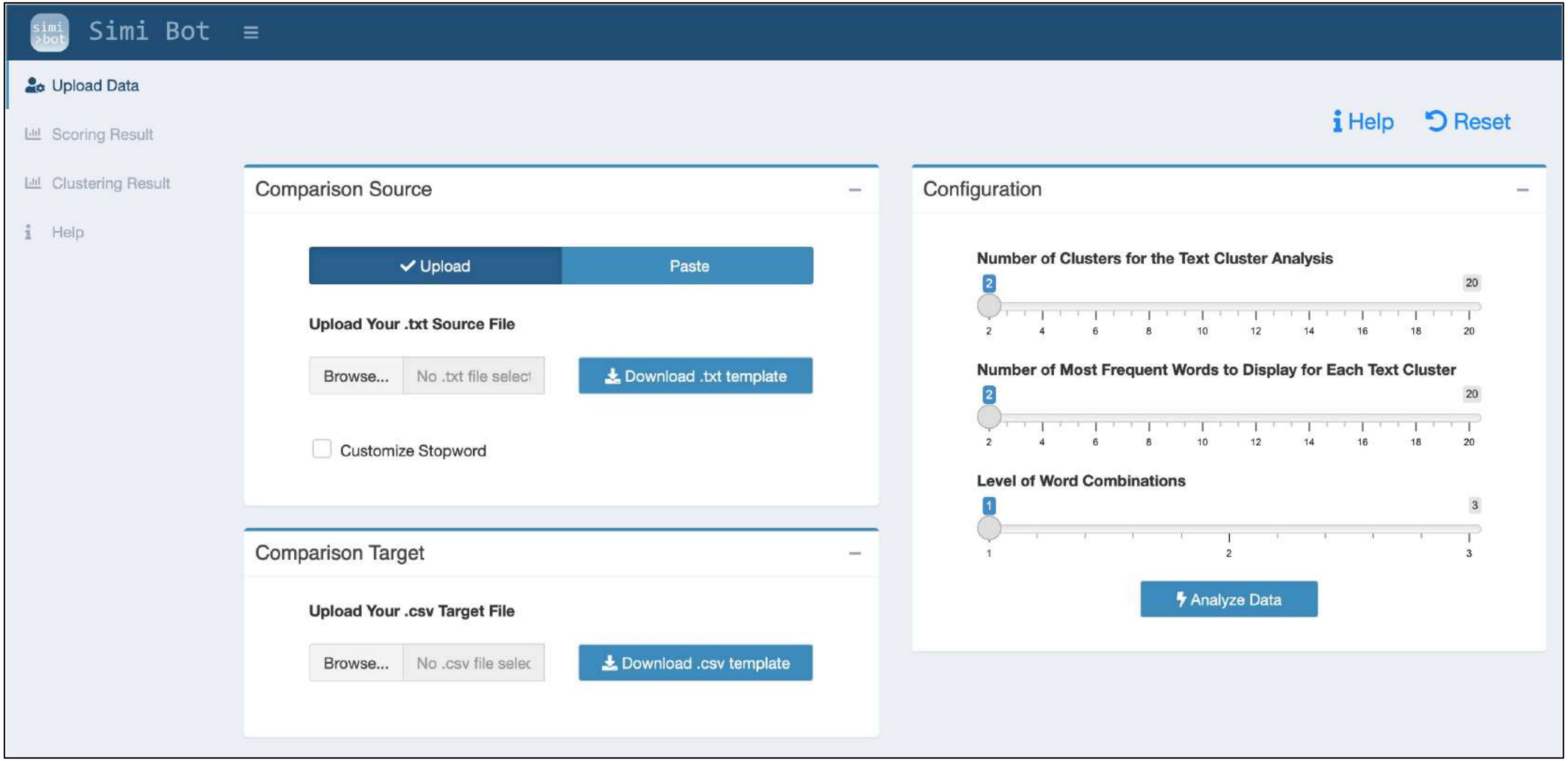
- We will use the tool Simi Bot for text similarity scoring
  - https://wukunchen.shinyapps.io/SimiBot/
- The results look like this:



TEXT	description	similarity_score
Text A	Most mornings I like to go out for a run.	0.099
Text C	The lead runner broke away from the pack early in the race.	0.091
Text B	Running is great exercsie for the brain.	0.083

- Let's try it

Simi Bot



Comparison Source

Upload Your .txt Source File

☐ Customize Stopword

Comparison Target

Upload Your .csv Target File

Configuration

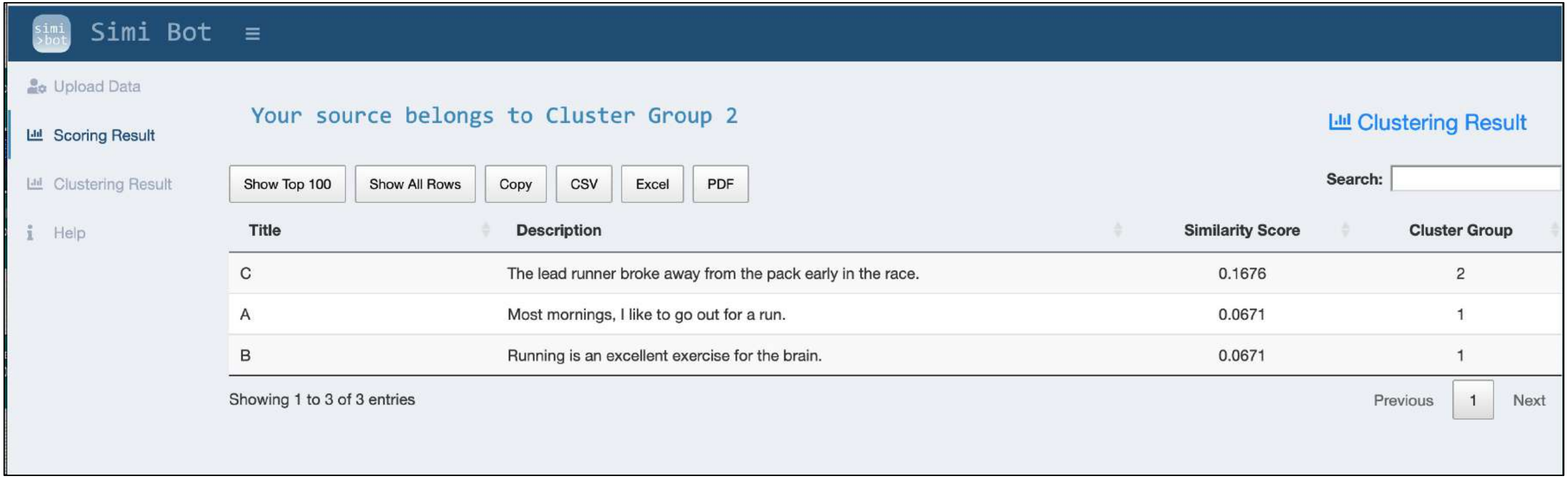
Number of Clusters for the Text Cluster Analysis: 2

Number of Most Frequent Words to Display for Each Text Cluster: 2

Level of Word Combinations: 1

FORTINO Global Education

Results

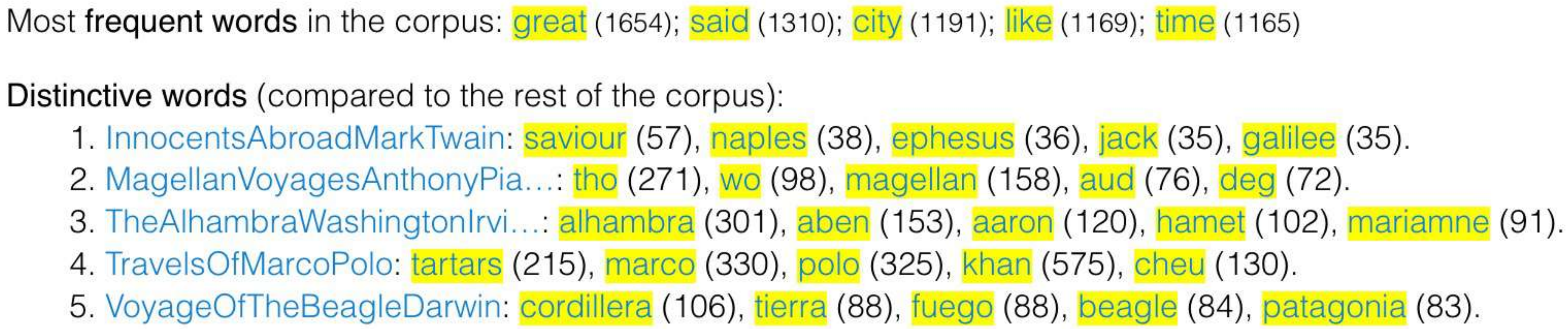


Your source belongs to Cluster Group 2

Title	Description	Similarity Score	Cluster Group
C	The lead runner broke away from the pack early in the race.	0.1676	2
A	Most mornings, I like to go out for a run.	0.0671	1
B	Running is an excellent exercise for the brain.	0.0671	1

Showing 1 to 3 of 3 entries

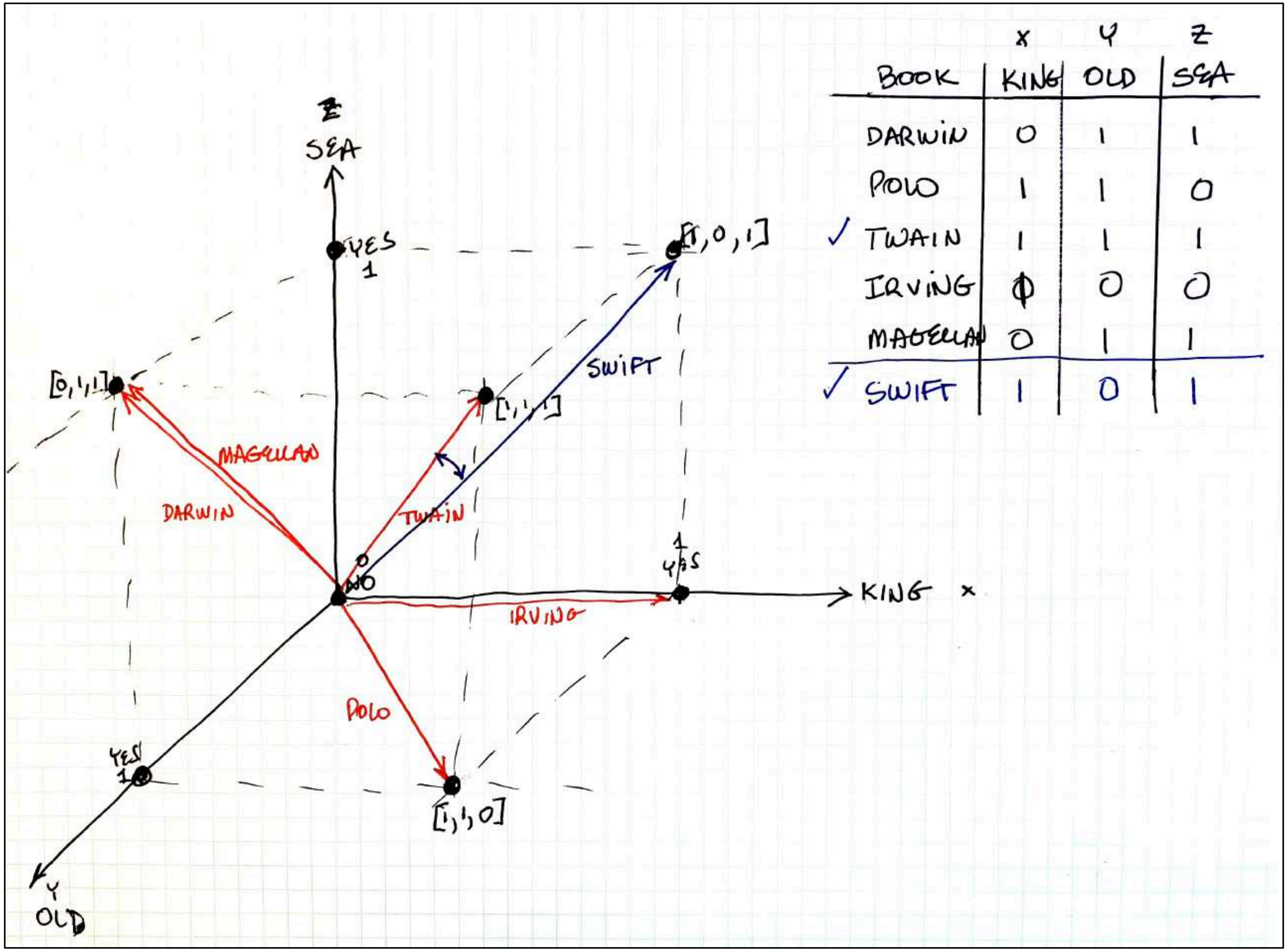
TF-IDF weighs distinctive words more



Most **frequent words** in the corpus: great (1654); said (1310); city (1191); like (1169); time (1165)

Distinctive words (compared to the rest of the corpus):

1. InnocentsAbroadMarkTwain: saviour (57), naples (38), ephesus (36), jack (35), galilee (35).
2. MagellanVoyagesAnthonyPia…: tho (271), wo (98), magellan (158), aud (76), deg (72).
3. TheAlhambraWashingtonIrvi…: alhambra (301), aben (153), aaron (120), hamet (102), mariamne (91).
4. TravelsOfMarcoPolo: tartars (215), marco (330), polo (325), khan (575), cheu (130).
5. VoyageOfTheBeagleDarwin: cordillera (106), tierra (88), fuego (88), beagle (84), patagonia (83).



	x	y	z
BOOK	KING	OLD	SEA
DARWIN	0	1	1
POLO	1	1	0
✓ TWAIN	1	1	1
IRVING	1	0	0
MAGELLAN	0	1	1
✓ SWIFT	1	0	1