Università degli Studi di Milano Bicocca | DISCO department | Data Science MSc
Exam project for the course ‘Data Semantics’

An analysis of the sacred texts of the 4 major religions using Word2vec embeddings

Yuliia Tsymbal, 894213
Tomasz Stanisław Grzesiak, 893734
Cosimo Simone Farallo, 889719
Diego Bartoli Geijo, 887208

Table of contents

- Introduction
- Preprocessing
- Training
- 3 research questions
- Future work
- References

Introduction

Word embeddings models aim to **quantify semantic similarities** between linguistic items based on their distributional properties in large samples of language data.

Distributional semantics research area studies the distributional properties of linguistic items. The basic idea of distributional semantics is that words that are used and occur in the **same contexts** tend to have **similar meanings**.

The **context of a word** is defined by its **nearest words**: *"a word is characterized by the company it keeps"* (John Rupert Firth, 1957).

Introduction - research questions

We will use **word embeddings** models to analyse the **sacred texts** of the **4 most popular religions**.

In particular we will **focus** on the following **questions**:

1. By exploring the **context** of some **interesting words**, is it possible to hypothesise certain **specific characteristics of a religion**?
2. By using embeddings trained on **aligned corpora,** is it possible to **link important characters** of different **religions**?
3. By exploring the **context** of words referred to **fundamental religious concepts** is it possible to find **emotional character associated with a religion**?

Introduction - corpora

The 4 most popular religions are **Christianism**, **Islam**, **Hinduism** and **Buddhism** (Wikipedia).

For each religion we consider the corresponding sacred text indicated by World Atlas.

To increase the size of the training corpora we will use **more translations** of **each** considered **sacred text** (Saeed et al., 2020). We decide how many translations to use based on the online availability for free and the number of total and unique words obtained.

We want to obtain **4 corpora**, one for each considered religion.

Preprocessing

We perform preprocessing with the following steps:

- convert **uppercase to lowercase**;
- remove **non alphanumeric** characters;
- remove **punctuation**;
- remove **stop words**.

Since we are not sure if performing lemmatization or not we save **lemmatized and non lemmatized version** for each corpora and compare them.

Notebook with complete preprocessing:
https://colab.research.google.com/drive/1bYwj7AUnUo7SO2nD3YKnLKik0h_h2I2D?usp=sharing .

Preprocessing - lemmatized corpora

Religion	**sacred text**	**N.of translations**	**N. words**	**N. unique words**
Christianity	Bible	3	998432	14544
Islam	Quran	8	784578	16735
Hinduism	Vedas and Upanishads	2 and 2	687782	15227
Buddhism	Tripitaka	5 books of the text	403179	15347

Preprocessing - non lemmatized corpora

Religion	**sacred text**	**N.of translations**	**N. words**	**N. unique words**
Christianity	Bible	3	1004875	19157
Islam	Quran	8	791238	22069
Hinduism	Vedas and Upanishads	2 and 2	697742	18874
Buddhism	Tripitaka	5 books of the text	404717	19699

Preprocessing - lemmatize or not

After a careful analysis, we decide **not to perform lemmatization** for 2 main reasons:

- it does **not** seem to **significantly improve** models stability;
- without lemmatization **words important for our analysis** have **higher frequency percentiles** which can increase their reliability (Hellrich&Hann, 2016).

Notebook with complete analysis:
https://colab.research.google.com/drive/1zOw09qbZC5UqvxITetvMkAkm2L3Tcz8F?usp=sharing .

Training - algorithm

We will create word embeddings using **Word2vec algorithm** in two different contexts:

- **without alignment** of different **corpora**;
- **with alignment** of **corpora** using **CADE**.

For the tasks that do not necessarily require alignment of corpora we will use both type of models and compare the results.

Training - stability problem

Due to their stochastic component **word embeddings** are **not stable**.

The **most similar words** to a given word **could change between models** even though the models are trained on the same corpora.

The stability problem **does not** have a **solution commonly** recognised as the **best** one.

Training - stability solutions

To **increase stability** of single embeddings we take the following decisions:

- we use **Skip gram method**, it works better on semantic tasks (Mikolov et al., 2013);
- we use Skip gram **negative sampling** with **5 noise words**, it seems to be more stable than Skip Gram Hierarchical Softmax (Hellrich&Hann, 2016);
- we perform **6 iterations over corpora** for each embedding, it seems to be a good trade off between computational cost and stability obtained (Hellrich&Hann, 2016)
- We set a **context window of 5**, a **minimum frequency of 10** and a **vector size of 300**, considering these to be commonly used values.

Training - stability solutions

To **increase** the **significance of our conclusions** we train **30 embeddings for each corpora** instead of one and we perform the analysis combining the results.

Hellrich and Hann (2016) cited this method as a possible solution to the stability problem but with **high computational costs**. In our case such costs are acceptable since the corpora are relatively small. The method has been used in an interesting work (Martina Schories, 2020) published on the blog Towards Data Science.

Notebook with complete training:
https://colab.research.google.com/drive/1vOkSaHwTEHVOY7-zpNN8_0Lc1vNRtPJp?usp=sharing .

First question

By exploring the context of some interesting words, is it possible to hypothesise certain specific characteristics of a religion?

We select 5 words that we consider interesting to explore: **woman**, **poor**, **wisdom**, **sins** and **death**.

For each of these words:

1. we search for the **5 most similar words** in **each corpora** (30 embeddings);
2. for each corpora, we report **all the words** appeared **grouped by frequency intervals** (5 most similar words change across 30 embeddings);
3. we **analyse** the **most frequent words** for each corpora and try to **draw some conclusions**.

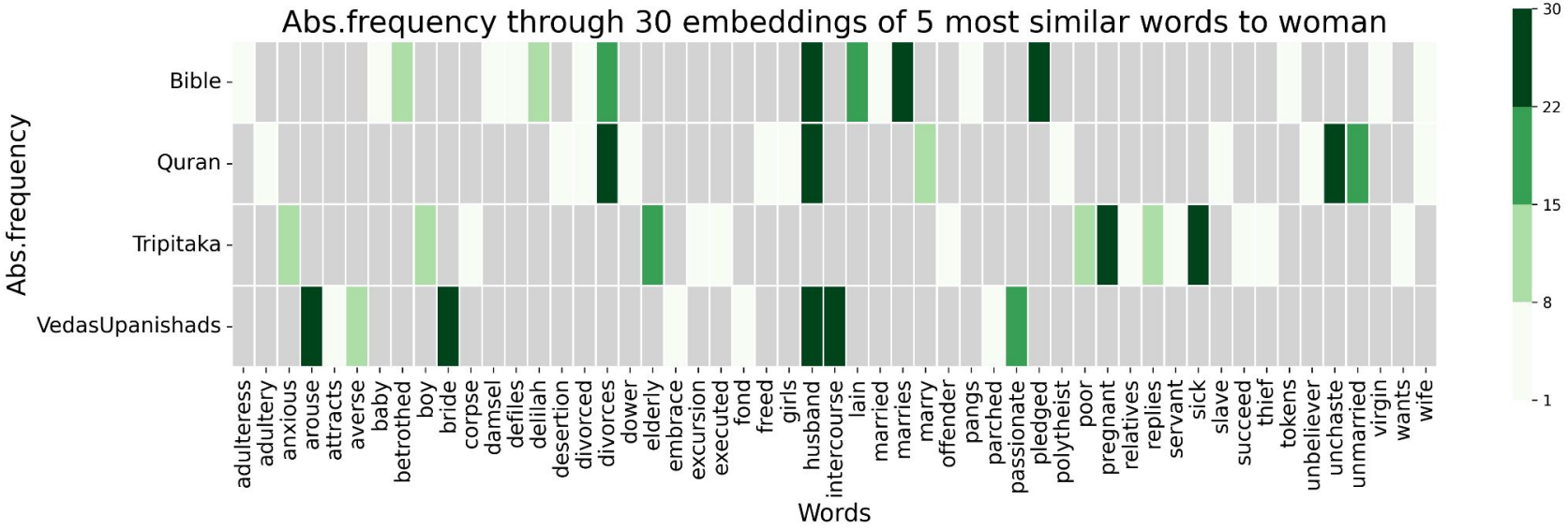
First question - remarks

We conducted this analyses both for models trained on non aligned corpora and on aligned corpora. We found the **results** with **models trained on non aligned corpora** to be **more interesting** and **easy to interpret** so we reported only these ones.

Words belonging to **frequency percentiles values between 90 and 99** tend to be **more stable** (Hellrich&Hann, 2016) . We check the frequency percentile of each word analyzed. We register **only two values** of frequency percentiles **lower than 90** for words studied: ‘poor’ for the Vedas and Upanishads and ‘sins’ for the Tripitaka.

Notebook with code:
https://colab.research.google.com/drive/15KdsNvu7hRS7cGCCLIFkkwGeKYNHxqk2?usp=sharing .

First question - ‘woman’



First question - ‘woman’

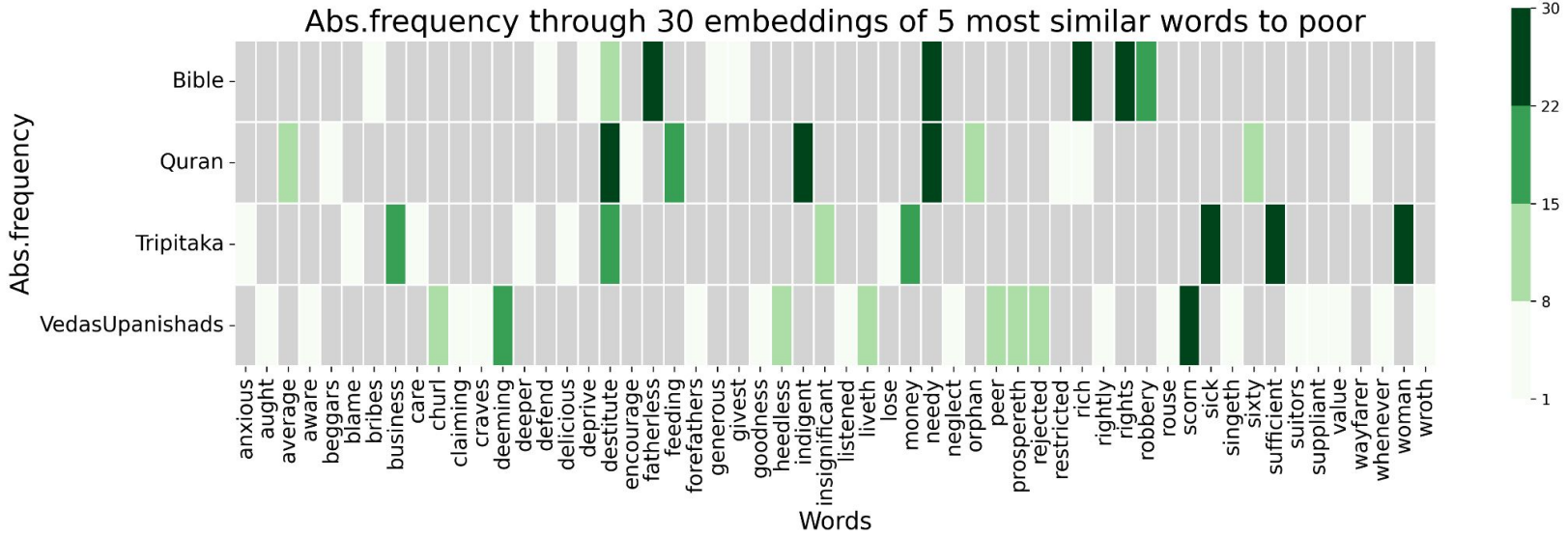
Bible: the most frequent words observed are husband, marries and pledged. The context of woman is **strongly related to** her role as **wife** and **bride**.

Quran: the most frequent words are husband, divorces and unchaste. The context of woman is **still related to** her role as **wife** but in a **more negative way**. The word unchaste introduces a **sexual aspect** not highlighted before.

Tripitaka: the most frequent words are pregnant, sick and elderly. The wife role is no more central while it acquires importance the role of the **woman as life bringer** and her **health conditions**.

Vedas and Upanishads: the most frequent words are husband, bride, intercourse and arouse. The context of woman is **still strongly related to** his **wife** and **bride** role but in a **more sexual way** than in other religions.

First question - ‘poor’



First question - ‘poor’

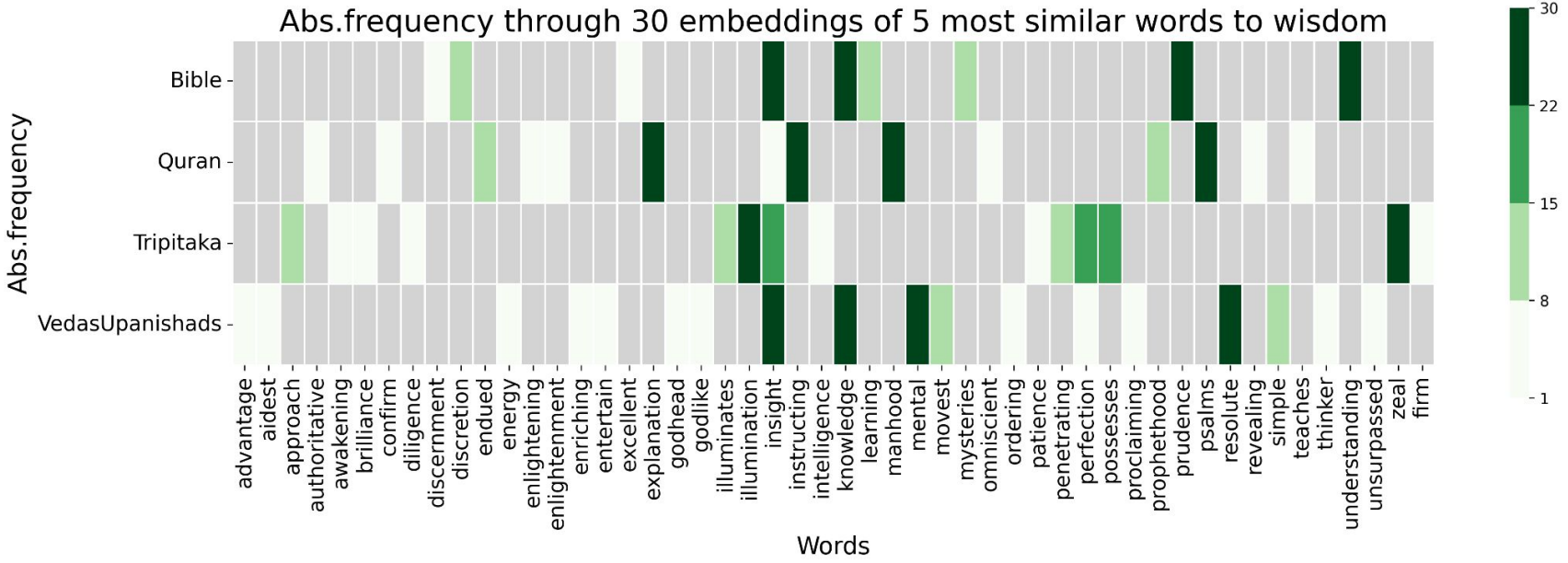
Bible: the most frequent words observed are fatherless, needy, rich and rigths. Two negative words and two positive ones, **poor** are **not despised**. The presence of the word rich highlight a fundamental concept of Christianity, **true richness** is **not material**.

Quran: the most frequent words are destitute, indigent and needy, general synonyms of poor. Here we don’t find **any positive words**.

Tripitaka: the most frequent words are sick, sufficient and woman, words difficult to interpret. Also money and business have high frequency. The poverty seems to be treated from a **mainly material point of view**.

Vedas and Upanishads: the most frequent word is scorn, reflecting a **harsh conception** of poor people. The words change a lot through the embeddings, the word poor seem to be less reliable in this embedding than in the others, probably due to the lower frequency of the word.

First question - ‘wisdom’



First question - ‘wisdom’

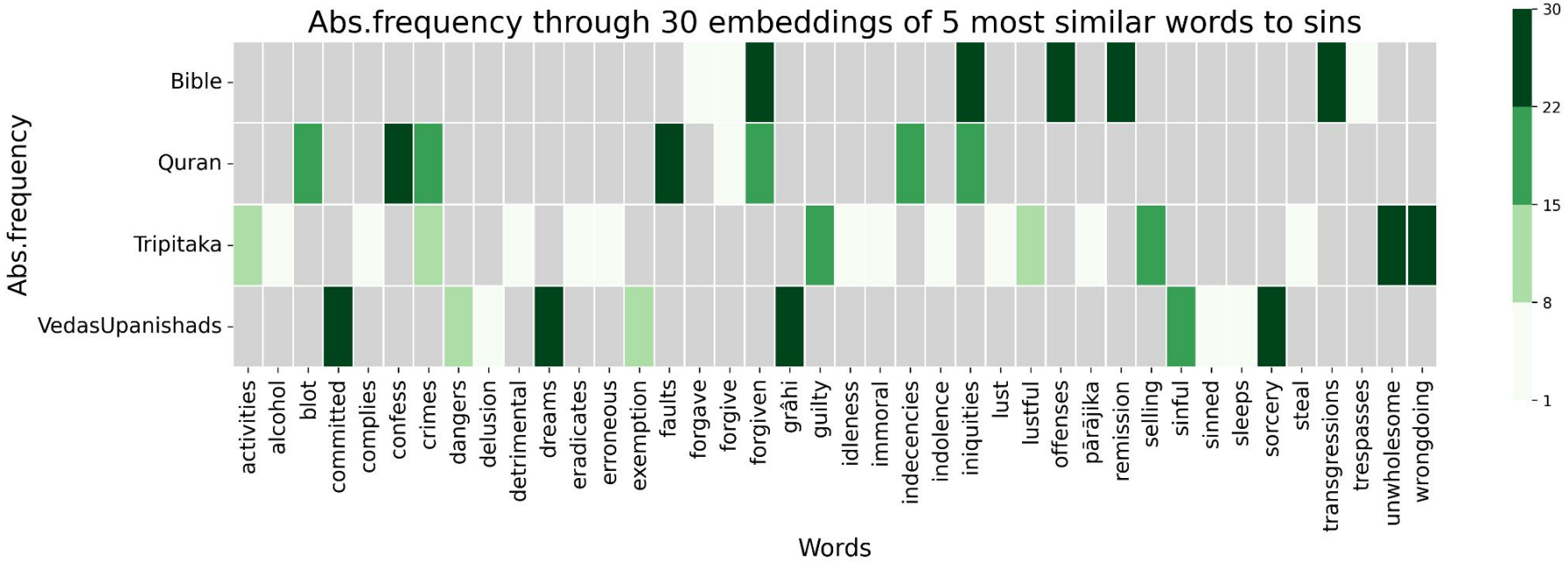
Bible: the most frequent words observed are insight, knowledge, **prudence** and **understanding**. The last two are particularly interesting because they appear only for the Bible.

Quran: the most frequent words are explanation, instructioning, manhood and psalms. Here the idea of wisdom is more ‘academic’, related with the **ability to teach** and with **books and scriptures**, and it’s **associated to men**.

Tripitaka: the most frequent words are illumination and zeal. The last one it is interesting because it suggests a more **‘practical’ idea of wisdom**.

Vedas and Upanishads: the most frequent word are insight, knowledge, mental and resolute. Wisdom seems to be considered a mix of **theoretical knowledge** and **ability to act**.

First question - ‘sins’



First question - ‘sins’

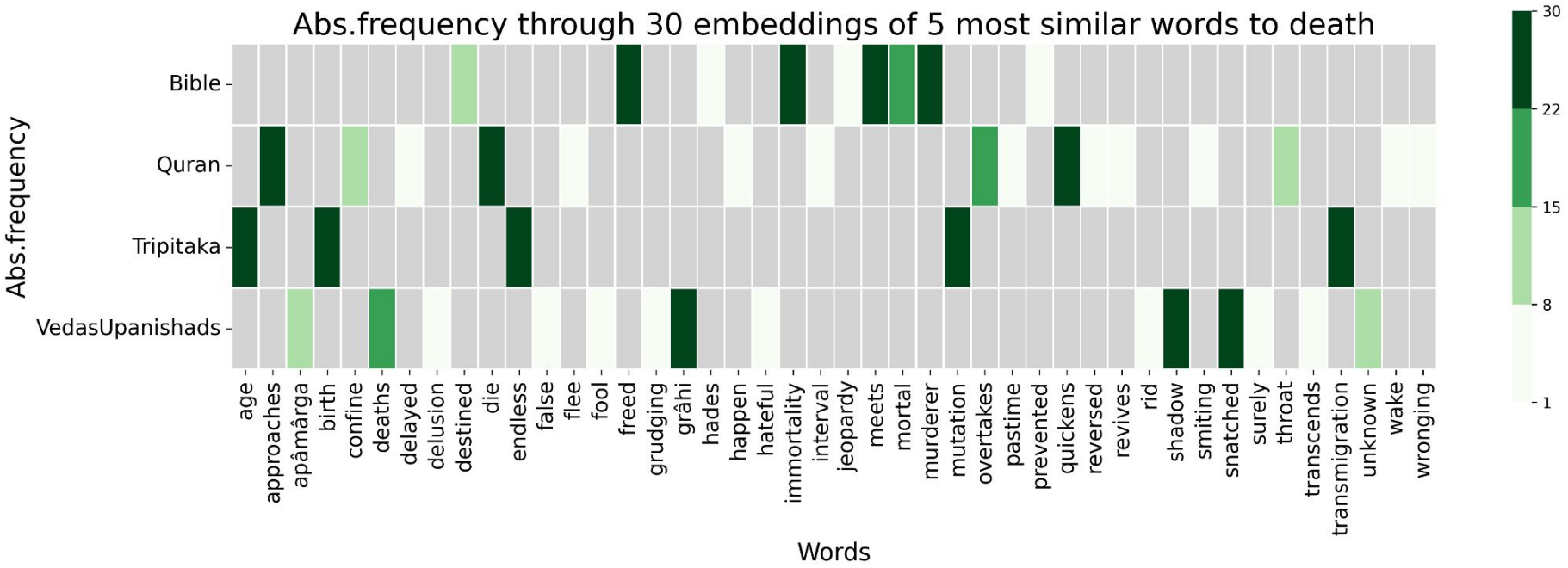
Bible: the most frequents words are forgiven, iniquities, offenses, remission, and transgressions. Although **3 of words are negative** the **other two** highlight the concept of **forgiven**, which is very important in Christianity.

Quran: the most frequents words are confess and faults. Forgiven, crimes or indecencies have also high frequencies. The **predominant aspect** is the **negative** one **but** also here are **present** the **forgiveness** and the **repentance**.

Tripitaka: The most frequents words are unwholesome and wrongdoing. In general all the words reflect an **exclusive negative** vision of sins.

Vedas and Upanishads:: the most frequents words are committed, grahi (they are demons?), **sorcery** and **dreams**. The last two words introduce a **‘fantastic’ meaning** to the word sins not present in the other texts.

First questions - ‘death’



First question - ‘death’

Bible: the most frequents words are freed, immortality, meets and murderer. The last two are difficult to interpret but the the first two are interesting because they suggest a **positive vision** of the death. In fact, Christianity sometimes describe death as a **liberation** of the soul that is immortal.

Quran: the most frequents words that are approaches, die and quickens. Also overtakes has high frequency. These words are not very clear but they seems to suggest a vision of the death as a **quick and violent natural process** that affects men.

Tripitaka: The most frequents words are age, birth, endless, mutation and transmigration. All the words refer back to the concept of **immortality**. **Death** is not seen as an end but **as a mutation, a new birth**. Worth to notice that the same 5 words appear for all the 30 models.

Vedas and Upanishads: The most frequents words are grahi, shadow and snatched. Grahi is a word used to identify bad health conditions, while shadow and snatched depict a **negative and violent vision of death**.

First question - answer

By exploring the context of some interesting words, is it possible to hypothesise certain specific characteristics of a religion?

The **conclusions** drawn with the analysis of the words reported are **quite interesting**. However we have also analysed a lot of words, not reported here, that led to meaningless results. Answering the first question we can conclude that the **task** is **possible** but **not linear**, it **requires some tries**.

Second question

By using embeddings trained on aligned corpora, is it possible to link important characters of different religions?

We select **4** words that refer to **important characters**, **one for** each **religion**: **jesus**(Christinaity), **abraham**(Islam), **vishnu**(Hinduism) and **buddha**(Buddhism).

We compute correspondences between spaces. For each of these words:

1. we search for the **10 most similar words** in **each corpora**(30 embeddings), excluding the one for which the word has been selected;
2. for each corpora, we report **all the words** appeared and we compute the **medium similarity** with the word studied across the 30 embeddings;
3. we **analyse** the **10 words with higher values of medium similarity** and look if among them we **find important characters** of the other religions .

Second question - remarks

We perform this analysis using models trained on **aligned corpora**.

We have initially selected the word Muhammad for Islam, but the results were more interesting for the word Abraham so we have decided to change.

Notebook with code:
https://colab.research.google.com/drive/15KdsNvu7hRS7cGCCLlFkkwGeKYNHxqk2?usp=sharing .

Second question - ‘jesus’

Quran

Words	Med.similarity
jesus	0.444103
effectual	0.394021
priority	0.391012
omit	0.386032
perceiving	0.383336
begged	0.381515
untrue	0.380948
vindicate	0.380548
joking	0.380518
encounters	0.380484

Tripitaka

Words	Med.similarity
talking	0.402373
disheartened	0.394797
pledging	0.387983
admitted	0.386864
sunakṣatra	0.379802
unwilling	0.379511
reputed	0.379269
heed	0.378623
derogatory	0.378491
udāyin	0.378347

Vedas and Upanishads

Words	Med.similarity
practicing	0.407554
buyer	0.395309
wretch	0.394719
bo	0.386422
describing	0.385509
repairing	0.382564
maturity	0.381507
mo	0.380219
dared	0.378399
expectation	0.377901

Second question - ‘jesus’

Most of the words aren’t personal names, not the desired outcome.

However, we found a link with some characters:

- Quran: **jesus**, he is present also in this text;
- Tripitaka: **sunaksatra**, the cousin of Buddha, and **udayin**, king of a region in India.

Second question - ‘abraham’

Bible

Words	Med.similarity
abraham	0.569496
isaac	0.408957
keturah	0.369329
epher	0.365060
honoring	0.362900
nahor	0.360912
obed	0.359453
respectfully	0.359300
jokshan	0.358895
sojourned	0.358619

Tripitaka

Words	Med.similarity
derogatory	0.364146
successors	0.355199
puṣkarasvādi	0.352717
brahmanical	0.352226
progeny	0.351696
rāhula	0.351383
whosoever	0.350793
kauṇḍinya	0.350662
kevaddha	0.350521
founder	0.350447

Vedas and Upanishads

Words	Med.similarity
truest	0.351577
generator	0.351465
maintaining	0.350602
avails	0.349353
dwellingplace	0.348112
tested	0.347390
rik	0.344817
rewarded	0.342427
purohita	0.341063
conducive	0.341040

Second question - ‘abraham’

Now we find some more personal names than before.

In particular we find the following links:

- Bible: **abraham**, he is also present in this text and other important prophets;
- Tripitaka: **rahula**, the only son of Buddha, and **kaundinya**, one of the first 5 buddhist monks;
- Vedas and Upanishads: worth to highlight the presence of the word **purohita**, the name of Hindu priests.

Second question - ‘buddha’

Bible

Words	Med.similarity
chieftain	0.392083
annoyed	0.375913
refusal	0.375832
prone	0.374954
contest	0.373272
harming	0.372608
traced	0.372344
expiated	0.370969
yearning	0.369866
attentively	0.369578

Quran

Words	Med.similarity
deuteronomy	0.380637
verbally	0.379003
redeemer	0.374461
sporting	0.368517
darling	0.367276
speculate	0.366247
clients	0.365983
birthright	0.364511
teacheth	0.363427
disassociated	0.362742

Vedas and Upanishads

Words	Med.similarity
conclude	0.367368
congregational	0.364759
extracted	0.363033
soldier	0.362145
rong	0.360639
arriving	0.360426
prostration	0.358996
heals	0.358750
earned	0.357291
edifying	0.357085

Second question - ‘buddha’

Most of the words that appear make sense but again we don’t find any personal names.

The only real interesting thing is that the most similar word to Buddha in the Bible is **chieftain**.

Second question - ‘vishnu’

Bible

Words	Med.similarity
chiefly	0.404035
hasting	0.383675
admit	0.382610
declined	0.382292
symbols	0.380229
approveth	0.380084
misuse	0.379234
defilement	0.378631
thinkest	0.378486
acceptest	0.377772

Quran

Words	Med.similarity
morality	0.422513
eminence	0.400231
justification	0.398781
timely	0.390383
blessedness	0.387871
shaping	0.380376
heroic	0.379411
ethics	0.378931
preserver	0.378428
asketh	0.378265

Tripitaka

Words	Med.similarity
narrated	0.376799
culmination	0.374540
sakṛdāgāmin	0.373104
equality	0.370977
accomplishing	0.370732
unconditioned	0.370684
dhyāna	0.370378
cosmic	0.370302
conventional	0.369526
semblance	0.369097

Second question - ‘vishnu’

For the Bible and the Quran we don’t find any personal names.

For the Tripitaka we still don’t find any personal name but we find some interesting words: **sakrdagamin**, used to identify an enlightened person, and **dyhana** which is a type of meditation.

Second question - answer

By using embeddings trained on aligned corpora, is it possible to link important characters of different religions?

Although we have found some links, the **results** are **not satisfactory**.

This could be due to multiple facts. Maybe the embeddings have **not** been trained with **enough corpora** to profitably perform a really difficult task such as **cross corpora analysis**. In addition, a better knowledge of the different holy texts would be useful to **better identify** which **characters** make sense to **try to link**, and which results we can expect. For example we get the most interesting results with the word Abraham, probably because the concept of ‘founder’, ‘forefather’ is a concept clearly present in all the 4 religions

Third question

By exploring the context of words referred to fundamental religious concepts is it possible to find emotional character associated with a religion?

Since emotional characters are often expressed in texts in form of **adjectives**, we will focus on them.

In order to extract an emotional character, we manually select 5 words that describe some general concepts in religion: **god**, **faith**, **heaven**, **hell** and **evil.**

For each of these words, we searched for synonyms within a sacred text by searching for the most similar words and using part of speech recognition to help filter out nouns. After picking best candidates for synonyms by hand, we then searched for adjectives most similar to these synonyms (and the original word), again, facilitation part of speech recognition.

Third question - remarks

Described exploration is conducted separately on each corpus (sacred text). Hence, technically it does not require for corpora to be aligned. However, we conducted this exploration twice - using models trained separately and aligned using CADE. We found models aligned with CADE to yield considerably better results than trained separately. For this reason, we did not include results obtained from not aligned models in this presentation.

Notebook with exploration based on not aligned models:
https://colab.research.google.com/drive/1KPLUgL7iLZ0Fv1GofBS5LnkMUqjJ7VOZ?usp=sharing

Notebook with exploration based on models aligned with CADE:
https://colab.research.google.com/drive/1-5nXM2f1th-K-wRxl22XPHzYpnQyblSk?usp=sharing

Third question - ‘god’

Religion	Synonyms	Adjectives (selection of most similar)
christianity	lord, thankful, bestowing	remarkable, undivided, charitable, comfortable, impartial, unlimited, unstained, untrue (?), sympathetic
islam	allah, dwelleth, allwise	unchecked, invincible, presumptuous, majestic, unconditioned, thoughtful, medical, majestic, wisest
buddhism	guise, śuddhāvāsa, gandharva, brahmakāyika	uttered, unconquerable, uninjured, uplifted, exemplary, coral
hinduism	savitr, savity, devapi	gradual, valid, mightest, largest, impervious, wisest, solitary, musical (!), honourable, neglectful (!)

What is worth noticing, out approach was able to find considerably less adjectives describing ‘god’ for buddhism than for other religions.

Third question - ‘faith’

Religion	Synonyms	Adjectives (selection of most similar)
christianity	virtue, patience, endurance	effectual, unspeakable, unlimited, justified, financial (?), incomparable, active, mutual, moral, favourable
islam	belief, lip, mettle	voluntary, nominal, efficacious, untroubled, induced, islamic, undaunted, unquenchable, courageous
buddhism	serenity, trust, dedication	unbounded, unreasonable, envious (?), unshakable, continual, unobstructed, eventual, defective (!), dogmatic
hinduism	noonday, purohita, medhyatithi	unfolded, unnecessary, valorous, neglectful (!), theological, attractive, auxiliary, spontaneous

Third question - ‘heaven’

Religion	Synonyms	Adjectives (selection of most similar)
christianity	sky, heavens, alms	remarkable, equitable, gigantic, enormous, visible, apprehensive, charitable, tremendous, comfortable
islam	sky, heavens, firmament	needful, unwearied, unrolled, unshakable, thunderous, loftiest, auxiliary, innumerable, astronomical
buddhism	earth, bṛhatphala, ābhāsvara, akaniṣṭha	uprooted, unite, uninterrupted, solid, untouched, incalculable, immaterial, final, precipitous
hinduism	earth, marvel, quickeneth, spheres	unrolled, lustrous (!), productive, supernatural, venerable, respectable, unsurpassable

Notice, for christianity and islam ‘heaven’ is associated mostly with ‘sky’, while for buddhism and hinduism it is rather associated with ‘earth’.

Third question - ‘hell’

Religion	Synonyms	Adjectives (selection of most similar)
christianity	gehenna, hades, sheol, damnation	criminal, gross, untouched, vicious, indescribable, indestructible, sudden, knowledgeable, irresistible
islam	roast, hellfire, destination, scorch	gigantic, envious, massive, final, unsettled, criminal, miserable, ruinous, unhappy, hypocritical, angered
buddhism	compression, torments, guards, avīci	black, gigantic, untouched, uninterrupted, induced, eventual, conscious, conspicuous, envious
hinduism	ownership, absence, diversity, impressions	unrolled, lustrous (!), productive, supernatural, venerable, respectable, unsurpassable

Similarly as in the case of ‘god’, we were able to find significantly less adjectives for buddhism than for other religions.

Third question - ‘evil’

Religion	Synonyms	Adjectives (selection of most similar)
christianity	suspicions, reform, disaster, perversity	mischievous, latest, lustful, courteous, unequal, sudden, lookest, unwholesome, untrue, worst, heinous
islam	exploit, hellfire, bids, temptations	bad, unwholesome, massive, induced, unwholesome, unreasonable, stupid, regrettable, ruinous, unhappy
buddhism	courses, blindness, deeds, pāpīyas	volitional, treacherous, desirous, conspicuous, coral (!), temporal, unlawful, meritorious, unconquerable, vicious
hinduism	dreams, foul, spirits, raksas	offensive, indigenous, unrighteous, voracious, uninterrupted, offensive, inauspicious, heinous, vicious

Third question - answer

By exploring the context of words referred to fundamental religious concepts is it possible to find emotional character associated with a religion?

Obtained results clearly show that our approach to the question yields meaningful results. For each religion we were able to find adjectives describing specific emotional character, consistent with our general knowledge of each religion.

What is more, considering each of the explored words separately, we can rewrite our question as '***How*** *the concept x (according to y religion) is?*'. Thanks to the approach taken by us, we can also utilize the intermediate step of analysis and using obtained synonyms to answer also the question: '***What*** *the concept x (according to religion y) is?*'. Both of these questions can be really interesting from the theological and ecumenical perspective.

Future work - possibilities

Ideas for expanding the work:

- try to use **different algorithms**, **methods** and **input parameters**;
- try to **increase** the **corpora**, using not only free copies;
- try to **improve** the **corpora**, maybe by **specific preprocessing operations** for each text based on prior knowledge of the texts;
- **align** separately **each pair of corpora** and analyse the corresponding results;
- perform **cross corpora** analysis not on single words but on **sets of words**, manually created or maybe by **clustering**.

References

S. Saeed, S. Haider and Q. Rajput, "On Finding Similar Verses from the sacred Quran using Word Embeddings," *2020 International Conference on Emerging Trends in Smart Technologies (ICETST)*, 2020, pp. 1-6, doi: 10.1109/ICETST49965.2020.9080691. https://ieeexplore.ieee.org/document/9080691

Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan. The COLING 2016 Organizing Committee.

Schories, M. (2020), *Using Word Embeddings for Journalistic Research*, Towards Data Science, https://towardsdatascience.com/using-word-embeddings-as-a-method-for-journalistic-research-ae82ffea7a62

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781., https://arxiv.org/abs/1301.3781