Relazione del progetto
“Text mining on sacred texts from various religions”

Indice:

1. Sentiment Analysis

Quando l'essere umano legge un testo, associa alle parole un significato, positivo o negativo, e l'insieme delle parole assume così un'unica accezione positiva o negativa. Questo concetto si può espandere aggiungendo altri livelli di positività o negatività, e anche altre emozioni.
La sentiment analysis cerca di raccogliere questa informazione dal testo.

Domande:

1. Quali testi sacri hanno un sentiment positivo e quali hanno un sentiment negativo?
2. Quali testi sacri sono i più positivi?
3. C'è correlazione tra il sentiment di un testo sacro e la quantità di fedeli?
4. Il sentiment della religione rispecchia le caratteristiche degli stati in cui va per la maggiore?

Analisi:

-- Antico Testamento --

La mia analisi è iniziata dalla Bibbia, ho preso il testo, la traduzione inglese di King James, contiene Antico e Nuovo Testamento, ciò mi permetterà poi di creare 2 dataset, uno per l'Antico e uno per il Nuovo Testamento, dividendo in due il file di testo.
King James Bible : https://www.gutenberg.org/files/10/10-0.txt

Per prima cosa ho messo in ordine il dataset, rimuovendo la parte iniziale che contiene l'introduzione, e la parte finale che contiene ringraziamenti vari e note.
Queste due parti erano ovviamente superflue ed inutili per la mia analisi.
Una volta fatto ciò, ho diviso ogni singola parola, e in una tabella vi ho associato il numero di occorrenze, in modo da poter analizzare solamente le parole più influenti all'interno del libro.

> Considerazione:
> Questo tipo di analisi non tiene conto dell'importanza che potrebbe avere una parola che appare anche una sola volta all'interno di un libro.

A questo punto ho rimosso i numeri in quanto non sono influenti per la mia analisi.

Ora ho 3 scelte per continuare con la mia sentiment analysis:

- Utilizzare il lexicon BING
- Utilizzare il lexicon AFINN
- Utilizzare il lexicon NRC

Ho deciso di affrontare tutte e 3 le situazioni.

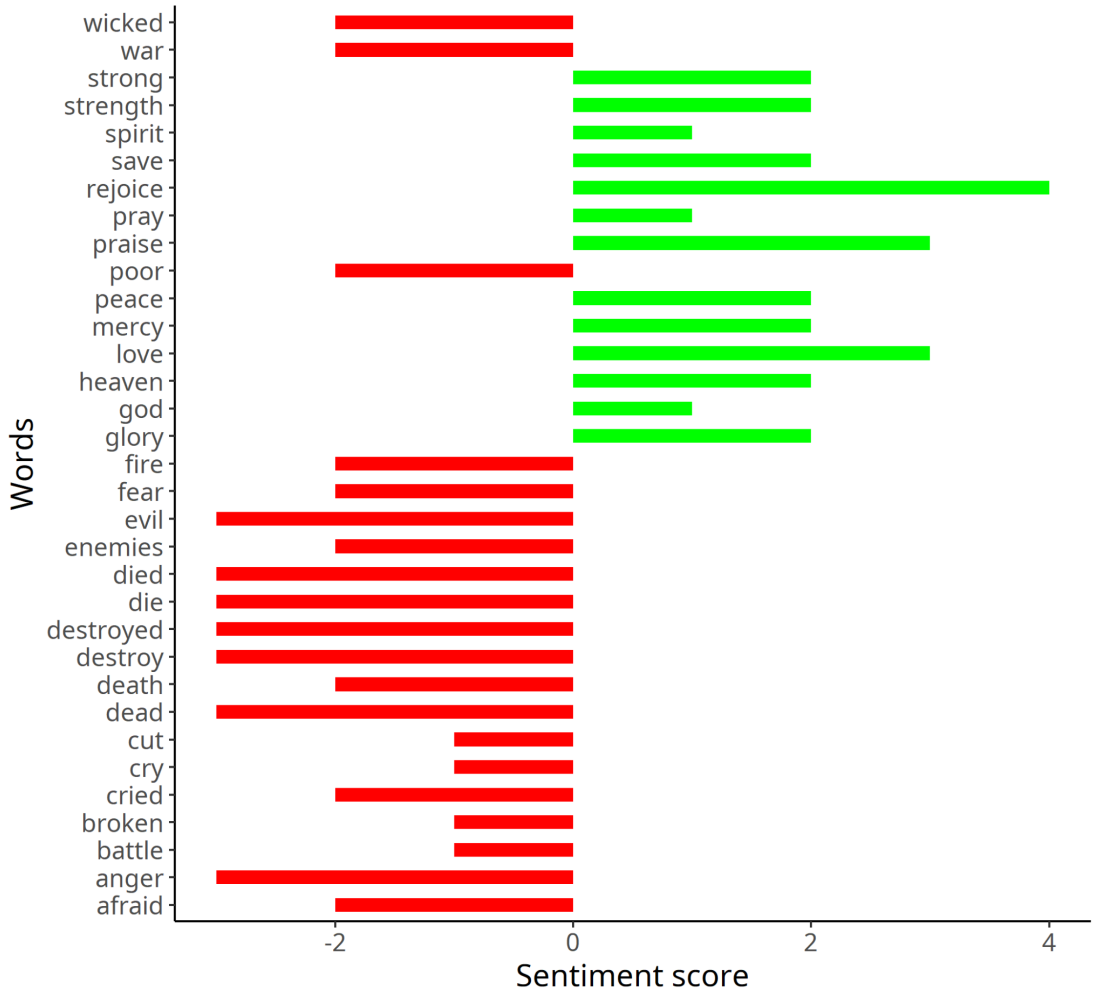
Ho creato 3 tabelle, effettuando un inner join con la tabella dei sentiment dei 3 lexicon che ho utilizzato, ho ottenuto come risultato 3 tabelle che contengono le parole dell'Antico Testamento, e:

- Nel caso di BING, un valore associato ad ogni parola tra *positive* e *negative*.
- Nel caso di AFINN, un valore *numerico* associato ad ogni parola, (la positività o meno del valore indica un certo livello di positività o negatività della parola).
- Nel caso di NRC, varie emozioni associate ad ogni parola.

Ora che ho 3 dataset su cui lavorare, decido come prima cosa di verificare la percentuale di parole con sentiment positivo o negativo all'interno dell'Antico Testamento.
Per fare ciò conto le righe della tabella creata dal lexicon BING che hanno un valore *negative*, la stessa cosa per le righe con valore *positive* (avrei potuto fare totale di righe - righe negative).
Una volta salvati questi 2 valori in delle variabili, applico la formula della percentuale, ed ottengo che:

- **Il 65.75% dell'Antico Testamento ha sentiment *negativo***
- **Il 34.25% dell'Antico Testamento ha sentiment *positivo***

Successivamente decido di plottare il livello di positività
(o negatività) che il lexicon AFINN associa alle parole dell'Antico Testamento.
Questo è il risultato:



Sono state plottate solamente le parole presenti più di 120 volte all'interno dell'Antico Testamento, sono circa una trentina e sono, a parer mio, le più rilevanti.

> Considerazione:
> Da questo grafico si evince che ci sono poche parole positive che appaiono tante volte e ciò è abbastanza in linea con il risultato precedente, che dava una percentuale maggiore di parole negative. Niente di sorprendente.

Infine, per osservare quali sono le principali emozioni che racchiudono le parole più frequenti dell'Antico Testamento, ho creato una wordcloud con le emozioni delle parole presenti più di 120 volte all'interno dell'Antico Testamento.

Come per con i risultati precedenti, non c'è un netto squilibrio tra emozioni positive e negative.

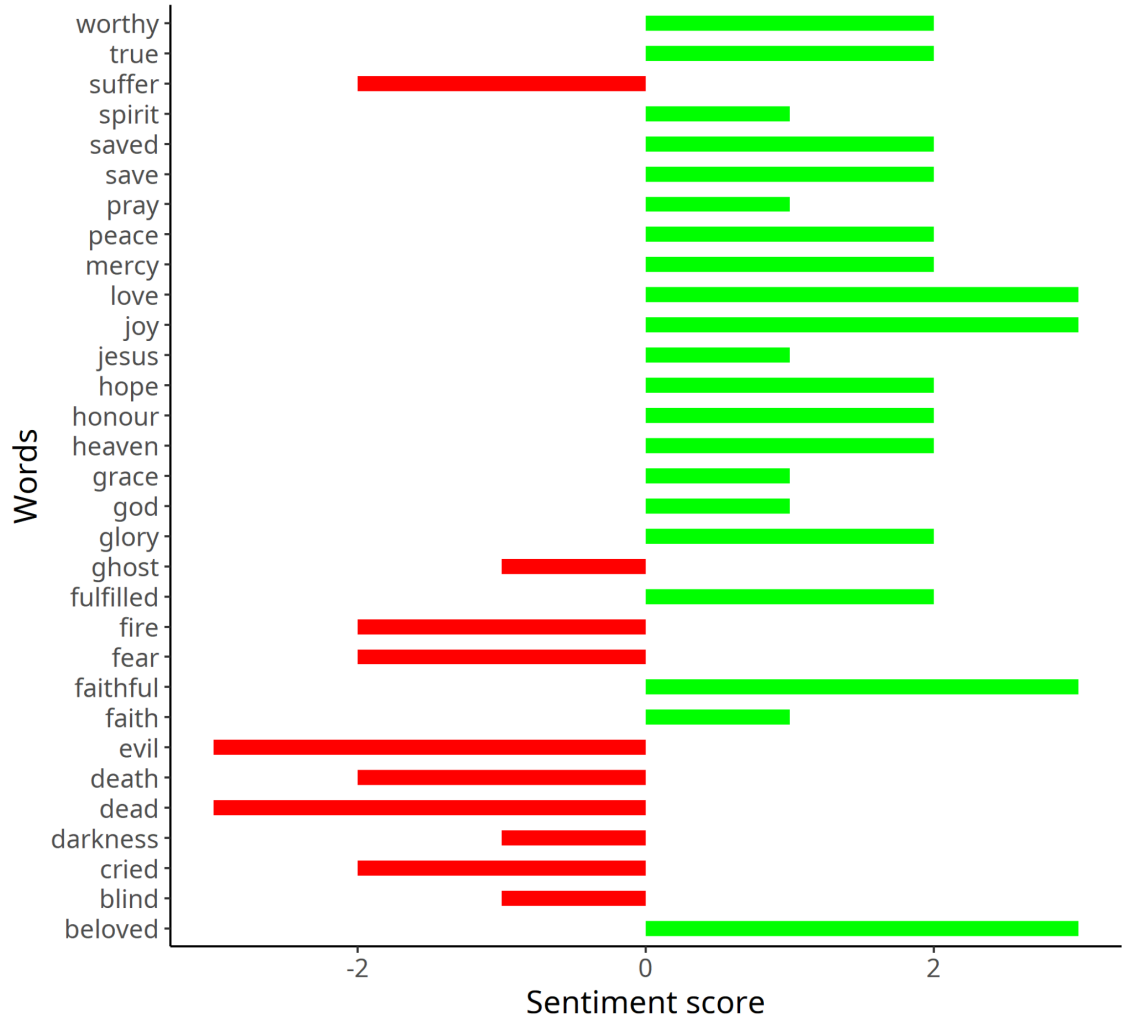
-- Nuovo Testamento --

Questo testo è stato ricavato dal King James Bible, dopodichè sono state fatte le stesse identiche operazioni precedenti.

Il dataset ricavato con BING, ha dato i seguenti risultati:

- **Il 63.79% del Nuovo Testamento ha sentiment *negativo***
- **Il 36.71% del Nuovo Testamento ha sentiment *positivo***

Il dataset ricavato con AFINN, ha prodotto il seguente grafico:



Sono state plottate solamente le parole presenti più di 50 volte all'interno dell'antico Testamento, sono anche qui circa una trentina e sono, secondo me, le più rilevanti.

Considerazione:

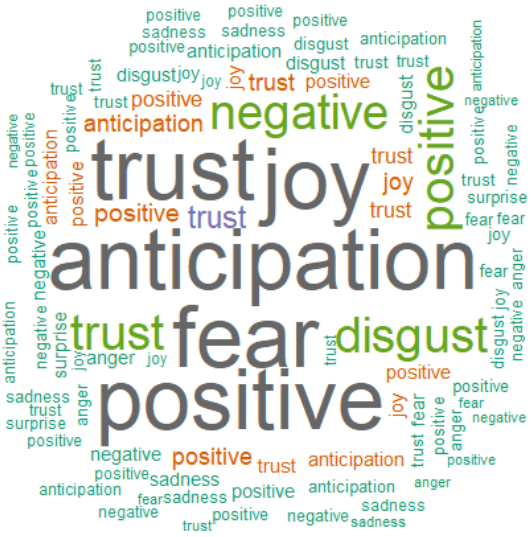
Da questo grafico si evince, al contrario dell'Antico Testamento, che ci sono molte parole positive che appaiono tante volte, nonostante la maggior parte delle parole del Nuovo Testamento siano negative.

Ciò significa che ci saranno tante parole negative che appaiono poche volte; questo fenomeno è curioso, e potrebbe indicare che in realtà a chi legge il Nuovo Testamento non sembra che sia poi così negativo, dato che le parole positive, seppur siano meno variate, appaiono più volte.

Considerazione:

Alcune parole non vengono contestualizzate effettuando questo tipo di analisi, come per esempio la parola "morte", a cui viene associato un sentiment negativo, ma all'interno del Nuovo Testamento viene utilizzata in contesti positivi, come per esempio quando si parla di Risurrezione.

Infine, la wordcloud ottenuta con le emozioni di NRC:



Anche qui, coerentemente con i dati precedenti, non c'è un netto distacco tra emozioni negative ed emozioni positive.

-- Corano --

Il Corano è stato preso sempre da Project Gutenberg, la traduzione inglese.

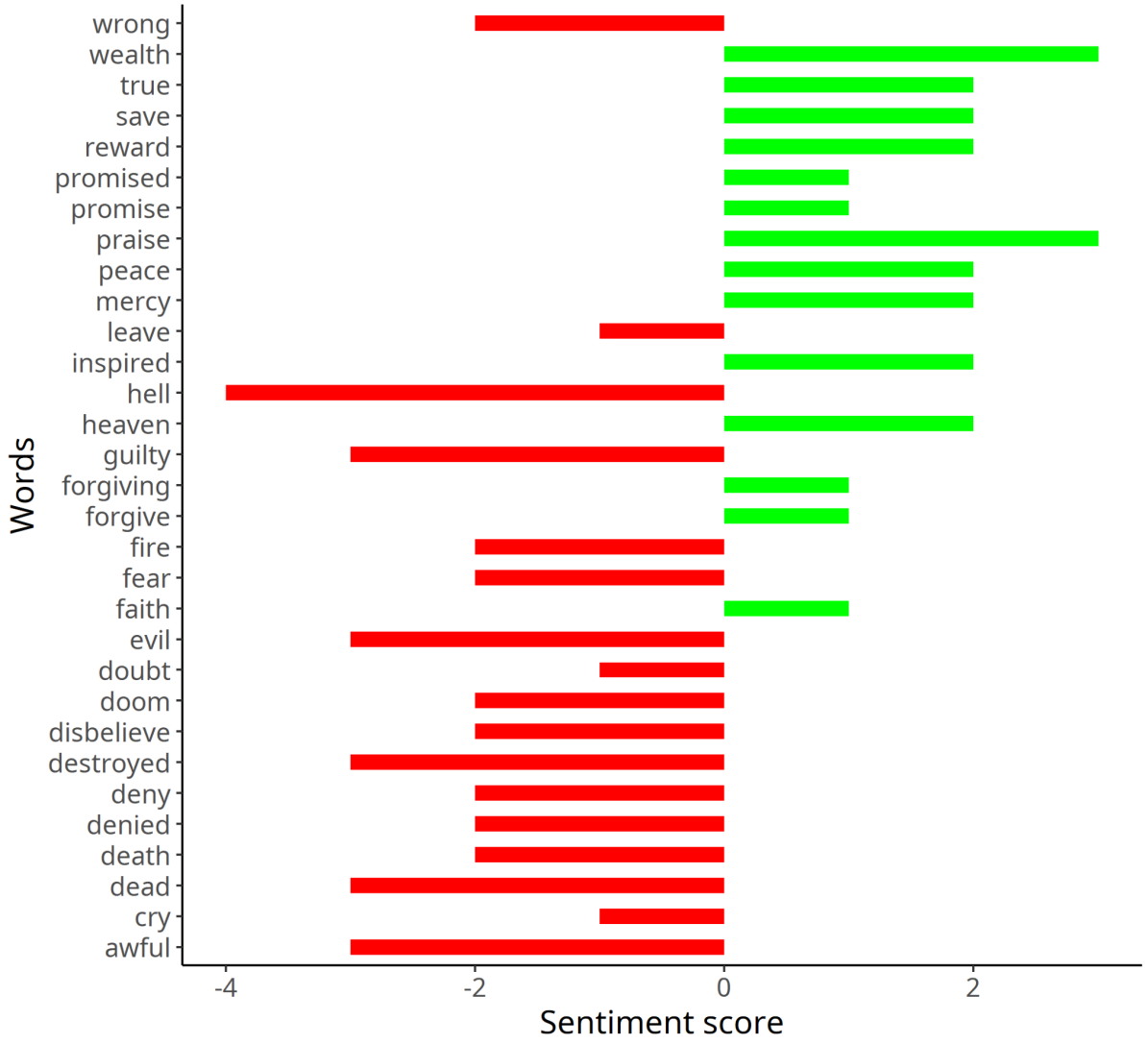
Quran: https://www.gutenberg.org/cache/epub/3434/pg3434.txt

Sono state effettuate le stesse identiche operazioni precedenti per l'ordinamento del dataset e risultati ottenuti sono i seguenti:

Secondo il lexicon BING:

- **Il 64.66% del Corano ha sentiment *negativo***
- **Il 35.34% del Corano ha sentiment *positivo***

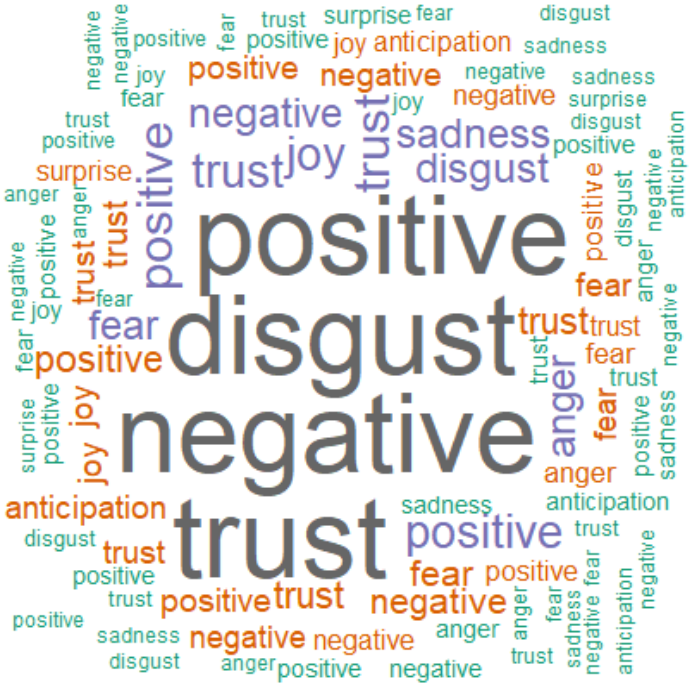
Con i dati ricavati grazie al lexicon AFINN, ottengo questo grafico:



Anche qui le parole considerate sono circa una trentina e sono quelle che appaiono più di 50 volte.

> Considerazione:
> Il grafico assomiglia molto a quello dell'Antico Testamento, anche i valori percentuali sembrano essere più vicini allo stesso. Mi aspettavo delle similarità tra queste 3 religioni, dato che venerano la stessa divinità, ma hanno profeti e riti diversi.

Infine, la wordcloud delle emozioni delle parole che appaiono più di 50 volte, secondo NRC:



Come in precedenza, non c'è un sostanziale squilibrio tra emozioni negative e positive che si noti d'impatto.

-- Rig Veda --

Ora esamino 3 religioni provenienti da tutt'altra parte, comincio con il Rig Veda, che è il testo sacro dell'Induismo.
Per questo testo, c'erano parti in lingua originale e parti tradotte.
https://www.gutenberg.org/cache/epub/14993/pg14993.txt

Come prima ho creato una versione tidy del mio dataset.

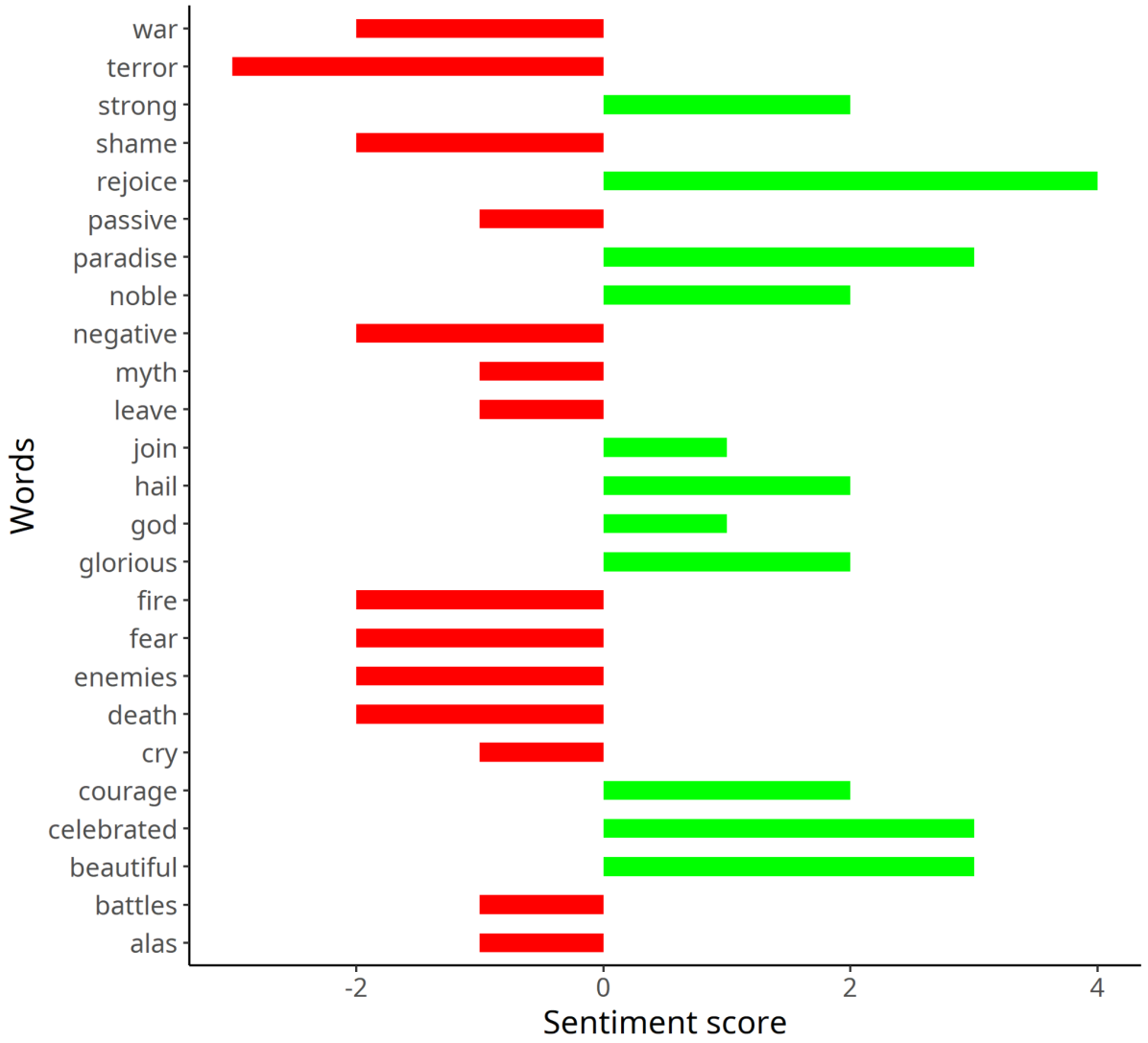
Il lexicon BING mi ha dato i seguenti risultati:

- **Il 51.88% del Rig Veda ha sentiment *negativo***
- **Il 48.12% del Rig Veda ha sentiment *positivo***

Considerazione:

Questa religione, sostanzialmente diversa dalle altre 3 che hanno una radice comune e sono anche abbastanza simili tra di loro, presenta un sentiment generalmente più positivo, e più equilibrato.

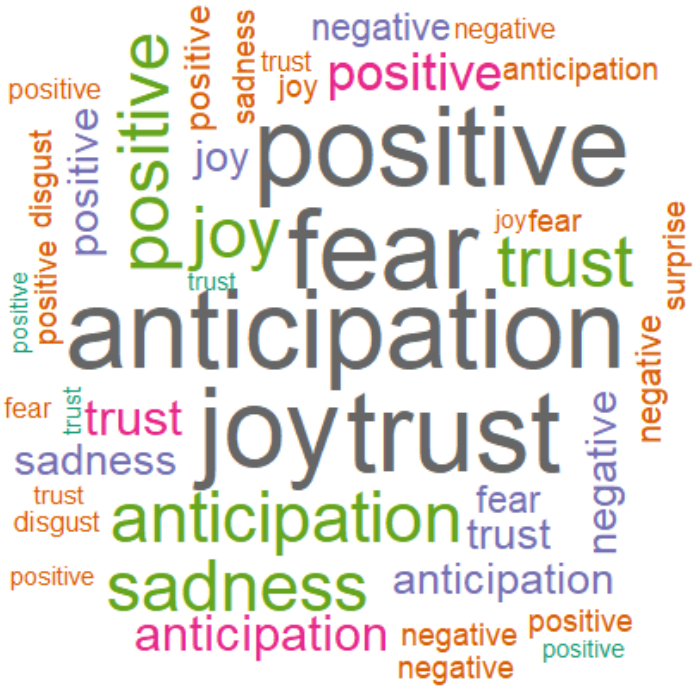
Grazie al lexicon AFINN, ho plottato questo grafico:



Le parole considerate, una trentina, sono quelle che appaiono più di 3 volte.

Considerazione:
In linea con i dati ricavati dal lexicon BING, c'è un sostanziale equilibrio tra positività e negatività osservabile anche qui.

Infine, la wordcloud con le emozioni delle parole più frequenti secondo NRC:



Come ci si aspettava, non si notano emozioni fortemente discrepanti d'impatto in questa wordcloud.

Considerazione:
Una cosa curiosa è che uno dei concetti e principi fondamentali dell'Induismo è la ricerca dell'equilibrio interiore ed il suo testo sacro principale è molto equilibrato dal punto di vista del sentiment.

-- Dhammapada --

È il testo sacro del Buddhismo e contiene parti in lingua originale e parti in inglese, ma le prime non verranno considerate.
https://www.gutenberg.org/files/35185/35185-0.txt

Come già fatto in precedenza, ho ordinato il dataset.

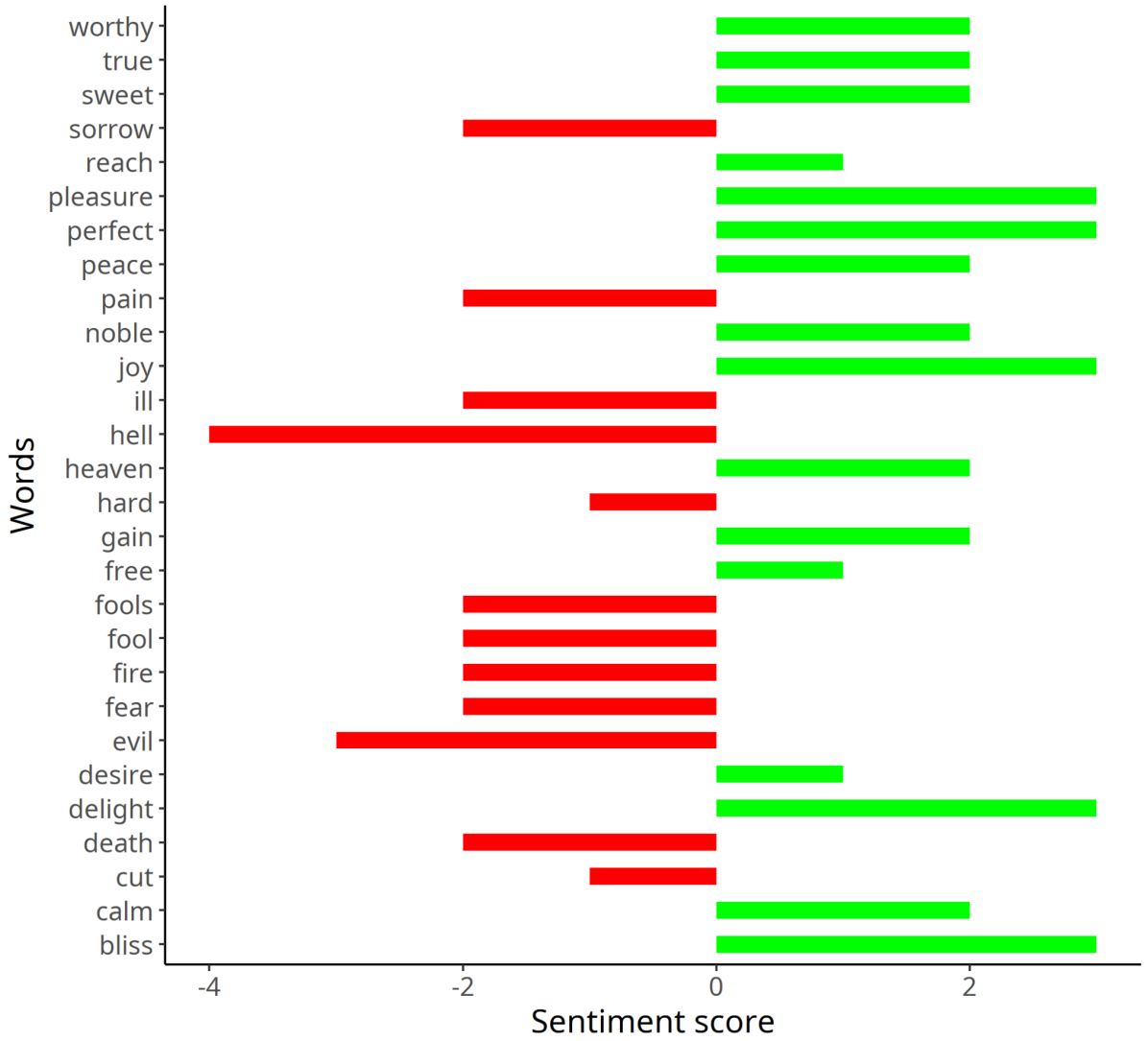
Dopodichè, il lexicon BING mi ha portato al seguente risultato:

- **Il 58.45% del Dhammapada ha sentiment *negativo***
- **Il 41.55% del Dhammapada ha sentiment *positivo***

Considerazione:

Anche qui il risultato differisce da quello ottenuto analizzando le 3 grandi religioni monoteiste e rappresenta un maggiore equilibrio, ma non tanto quanto nel caso del Rig Veda.

Il lexicon AFINN mi ha aiutato ad ottenere il seguente grafico:



Le parole considerate sono una trentina, e sono quelle che appaiono più di 8 volte.

Considerazione:
Coerentemente al risultato ottenuto con il lexicon BING, questo grafico ci mostra un certo equilibrio tra positività e negatività.

Considerazione:
Sembra essere leggermente squilibrato verso il positivo,
mentre nella complessità del libro il sentiment è leggermente squilibrato verso la negatività, quindi sembra essere lo stesso effetto di cui soffre il Nuovo Testamento, in cui le parole negative sono di più ma appaiono meno volte.
La curiosità è che il Buddhismo è nato dall'Induismo, come corrente alternativa e grazie a Buddha… allo stesso modo il Cristianesimo si potrebbe dire che è nato come corrente alternativa dell'Ebraismo, grazie a Gesù Cristo… di conseguenza queste due religioni presentano una caratteristica comune molto curiosa.

Per quanto riguarda la wordcloud delle emozioni secondo NRC:

-- Shri Guru Granth Sahib Ji (Adi Granth) --

È il testo sacro del Sikhismo, come per i 2 precedenti contiene parti in lingua originale e parti in inglese.

Link molto lungo

Dopo aver ordinato il dataset, ho proceduto con la mia analisi.

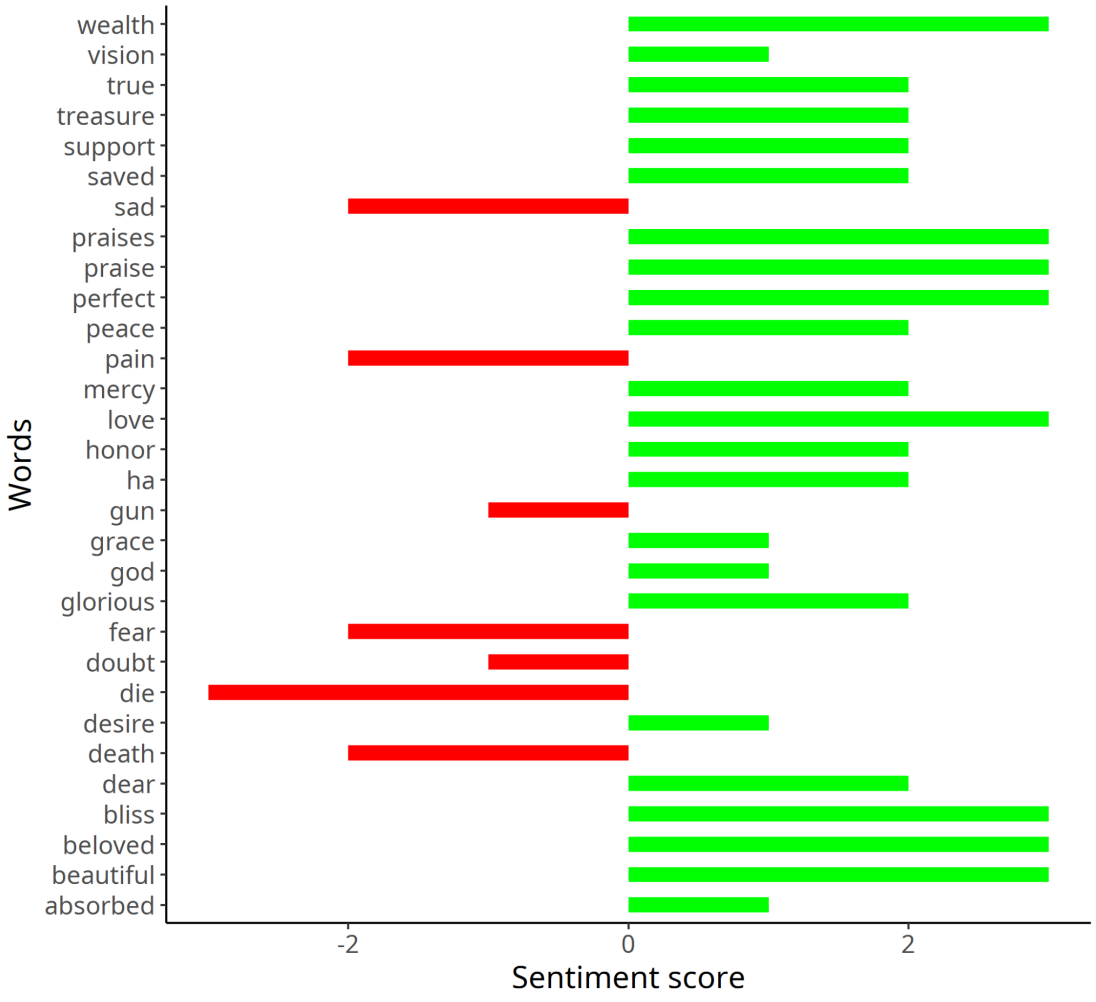
Il lexicon BING mi ha portato a queste conclusioni:

- **Il 63.25% dell'Adi Granth ha sentiment *negativo***
- **Il 36.75% dell'Adi Granth ha sentiment *positivo***

Considerazione:

Una curiosità che emerge, è che tutte le religioni monoteiste hanno un sentiment perlopiù negativo, anche quelle provenienti da culture molto distanti, infatti il sentiment dell'Adi Granth assomiglia molto al sentiment dei testi sacri delle 3 grandi religioni monoteiste, ed il Sikhismo è una religione monoteista.

Il lexicon AFINN mi ha portato ad elaborare questo grafico:



Considerazione:

Similarmente al grafico del Nuovo Testamento, sembra che le parole positive appaiano tante volte e che quelle negative appaiano poche volte.

Infine, la wordcloud con le emozioni ricavate grazie al lexicon NRC:

Al solito, seppure ci sia uno squilibrio verso la negatività, questo non è così estremo ed evidente.

Conclusioni:

- Ho usato il lexicon BING per osservare quante parole positive e quante parole negative sono presenti, non in quantità ma in varietà.
- Ho usato il lexicon AFINN per plottare le parole più frequenti, rappresentandone il grado di positività o negatività.
- Grazie all'analisi "particolare" fatta con BING, ciò mi ha portato a conclusioni particolarmente interessanti:
    - I testi delle 3 grandi religioni monoteiste hanno un sentiment simile e leggermente tendente al negativo

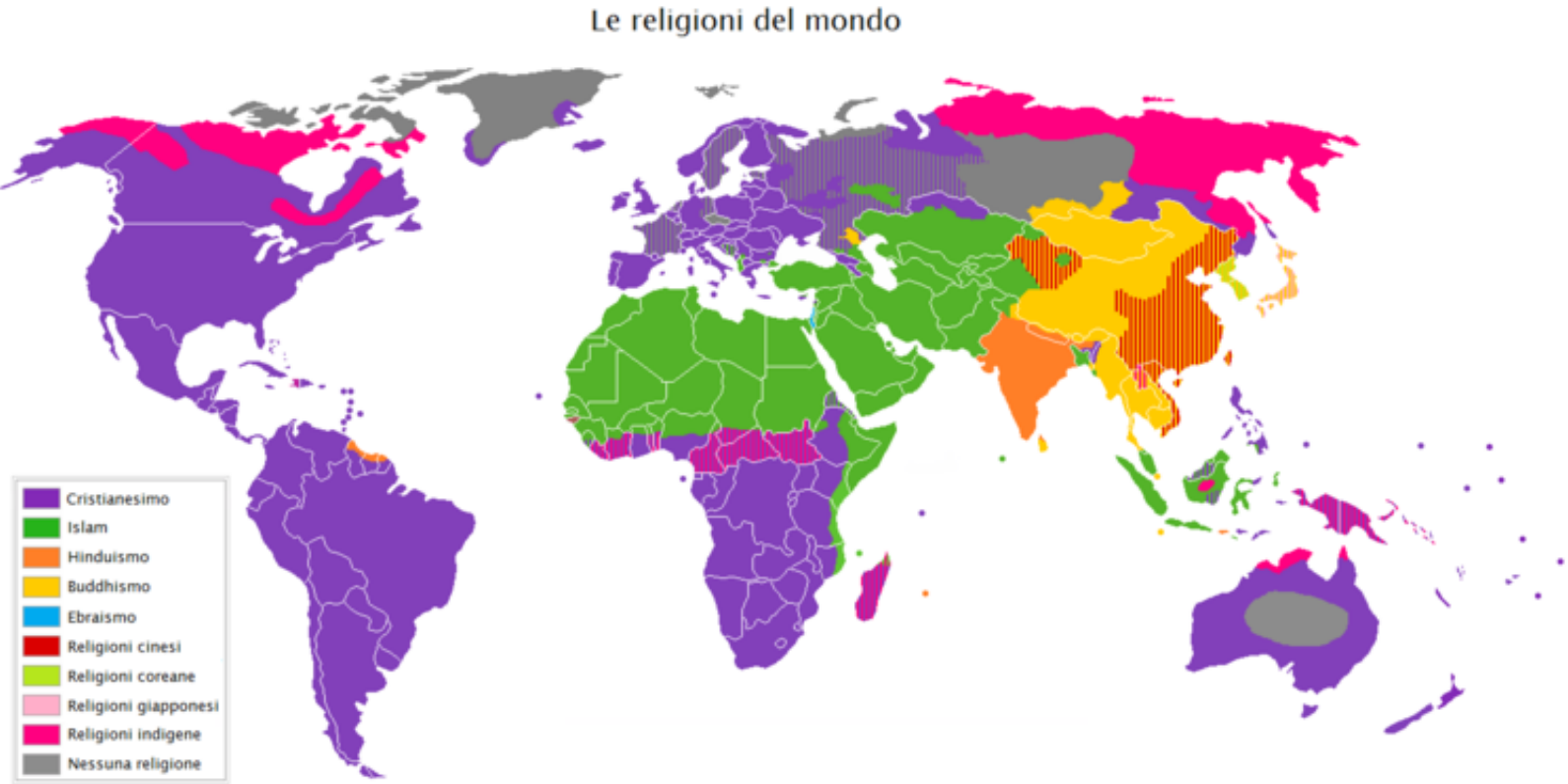
        - Il Nuovo Testamento contiene molte parole negative diverse, ma quelle positive sono presenti in frequenza maggiore.
    - I testi di Induismo e Buddhismo sono più equilibrati e simili
        - Il Buddhismo è una religione che deriva dall'Induismo, come una corrente alternativa.
          Allo stesso modo il Cristianesimo con l'Ebraismo.
            - Entrambi i testi sacri di queste due religioni sembrano avere un sentiment leggermente negativo, ma le parole positive sono presenti più spesso di quelle negative.
              Questa caratteristica lega le due religioni.
        - Uno dei principi cardine dell'induismo è la ricerca dell'equilibrio interiore ed il testo sacro dell'Induismo è esso stesso molto equilibrato.
    - Il testo del Sikhismo ha un sentiment molto simile a quello dei testi delle 3 grandi religioni monoteiste ed il Sikhismo stesso è una religione monoteista.

Risposte:

1. Quali testi sacri hanno un sentiment positivo e quali hanno un sentiment negativo?
    - I testi delle religioni monoteiste (Ebraismo, Cristianesimo, Islam, Sikhismo) hanno un sentiment leggermente negativo.
    - I testi delle religioni Induismo e Buddhismo sono leggermente più equilibrati.
2. Quali testi sacri sono i più positivi?
   Il testo sacro più positivo è quello dell'Induismo, il Rig Veda.
3. C'è correlazione tra il sentiment di un testo sacro e la quantità di fedeli?
   Ebraismo: 14 milioni

Cristianesimo: 2.2 miliardi
Islam: 1.8 miliardi
Induismo: 1.1 miliardi
Buddhismo: 1.5 miliardi
Sikhismo: 28 milioni
Sostanzialmente non c'è nessuna correlazione tra il numero di fedeli e la positività del testo sacro della religione di appartenenza.

4. Il sentiment della religione rispecchia le caratteristiche degli stati in cui va per la maggiore?



Non c'è un legame particolare tra la qualità della vita in uno stato ed il sentiment del testo sacro della religione che in quello stato va per la maggiore, anche se Cristianesimo ed Ebraismo, che hanno i testi sacri con sentiment più negativi, sembrano andare per la maggiore negli stati in cui mediamente la popolazione è più ricca, ovvero i paesi occidentali, mentre le altre religioni sembrano andare per la maggiore in stati emergenti o poveri.

L’analisi andrebbe approfondita con uno studio del PIL, e del numero più preciso di fedeli, ma non è questo l’intento del mio progetto.

2. Word and Document Term Frequency

Analizzando la frequenza con cui le parole si ripetono all'interno di uno o più testi si possono dedurre diverse informazioni, per esempio se le stesse parole sono utilizzate in 2 o più testi, si possono classificare le parole più usate nei testi presi in considerazione…
Questa analisi può essere fatta anche considerando i cosiddetti n-grams (così come la sentiment analysis), ma io non l'ho fatto, e in seguito spiegherò il perchè.

Domande:

1. Quali sono le parole più ricorrenti all'interno dei testi sacri?
2. Testi sacri di religioni molto differenti, hanno parole in comune?
3. In caso affermativo, può essere dovuto ad una qualche origine in comune tra le religioni, come la posizione geografica?

Analisi:

--3 Grandi Religioni Monoteiste--

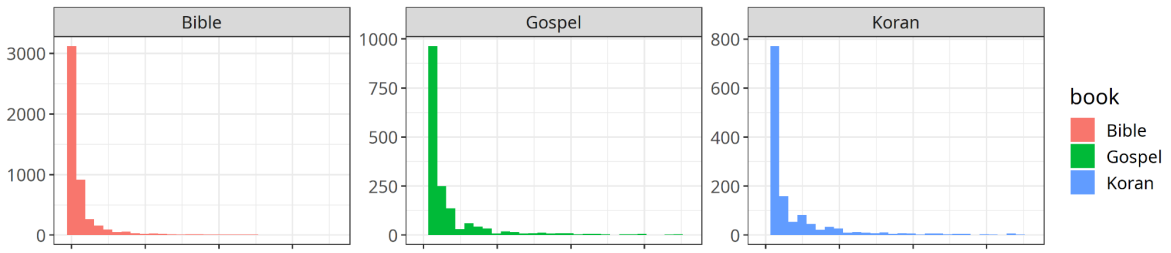
Per effettuare questa parte del progetto, ho considerato più libri insieme in 3 raggruppamenti, nel primo:

- Antico Testamento (Bibbia)
- Nuovo testamento (Vangelo)
- Corano

Come al solito ho messo in ordine i dataset ed ho creato una tabella contenente le parole e la loro frequenza assoluta. Alcune stop words le ho rimosse manualmente in quanto, facendo parte di un inglese arcaico (ye, thou, thine, …), non sono presenti nel dataset delle stop words.

Dopodichè ho messo insieme tutti i dataset con un full_join ed ho inserito una colonna per contenere il libro in questione.

Questa prima analisi non considera le parole in comune, ma tutte quante le parole.
Con la tabella ottenuta ho plottato la distribuzione di frequenza delle parole all'interno dei 3 testi:

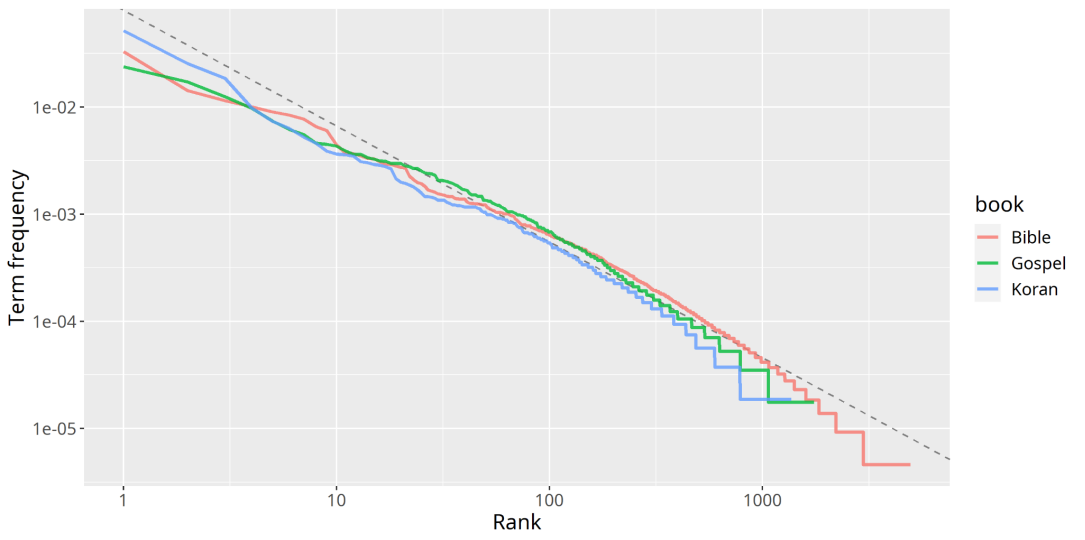


Word distribution among the
3 biggest monotheist religion's sacred texts

Come si può notare c'è una distribuzione molto comune, poche parole che appaiono tante volte e molte parole che appaiono poche volte. Questa relazione è chiamata legge di Zipf.
Per confermarlo ho deciso di sfruttare questa informazione generando un ranking per le parole:

- Ho stabilito un rango per ogni parola in ordine inverso rispetto alla frequenza, l'ho fatto ordinando il dataset in ordine decrescente rispetto alla frequenza e poi ho aggiunto una colonna con il numero di riga per il rango.
- Secondo la legge di Zipf, il rango è inversamente proporzionale alla frequenza.

Ottengo una tabella contenente la frequenza di occorrenza delle parole e un ranking associato ad ogni parola, che mi permette di plottare un grafico del rango contro la frequenza:



La scala utilizzata è $\log_{10}$.

Considerazione:

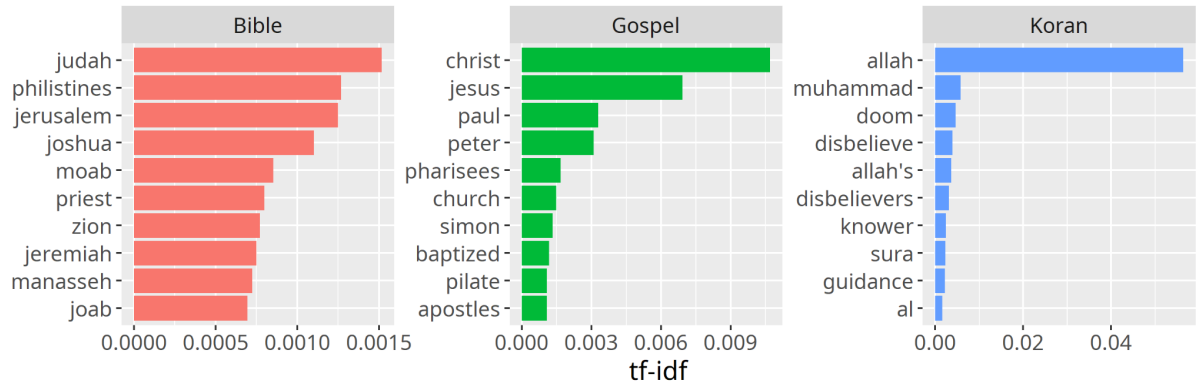
Il grafico ricorda molto una distribuzione classica della legge di Zipf, ciò indica che la distribuzione di frequenza è approssimativamente esponenziale.

Poche parole sono ripetute moltissime volte, per tutti e 3 i testi, e molte sono ripetute poche volte.

Le deviazioni ai bassi ranghi indicano che nei testi sacri c'è un uso minore delle parole più comuni del linguaggio e questo probabilmente è dovuto all'antichità dei testi.

Mi incuriosisce il fatto che dei testi così antichi, rispettino la legge di Zipf come i testi moderni, però probabilmente ciò è dovuto al fatto che questi testi sono stati tradotti e reinterpretati più volte nel corso dei secoli.

Ora analizzo la statistica tf-idf dei 3 testi sacri delle 3 grandi religioni monoteiste, il grafico ottenuto è il seguente:



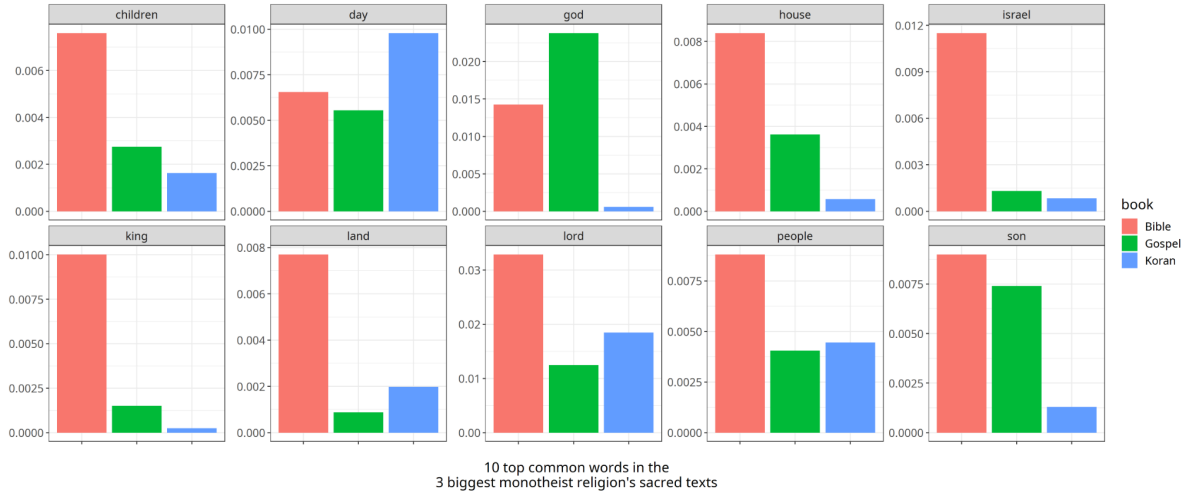
Considerazione:

Nel Corano la parola "Allah" è estremamente presente, ad indicare probabilmente la grande devozione che i fedeli mussulmani hanno nei confronti del loro dio.

Subito dopo troviamo la parola "Maometto", in questo testo il Dio ed il profeta sono evidentemente molto più importanti del resto.

Per quanto riguarda il Nuovo Testamento, la parola “Dio” non è tra le più frequenti, mentre “Gesù”, “Cristo”, i Santi “Pietro” e “Paolo” sono molto ripetute, ad indicare che anche qui i profeti della religione assumono una grandissima importanza, diversamente dal Corano, dove il dio acquista un’importanza maggiore.
Anche “Pilato” viene annunciato diverse volte all’interno del testo, così come i nomi degli apostoli. Sembra appunto che sia una religione in cui la devozione è molto rivolta alle persone che l’hanno creata.
Infine per ciò che concerne la Bibbia, le parole più frequenti sono “Giuda”, una delle dodici tribù d’Israele di cui si parla nell’Antico Testamento, la più grande ed importante, i “Filistei”, che erano un popolo localizzato nella Palestina.
“Gerusalemme” ha una grande importanza, più che nel Nuovo Testamento, “Moab”, una regione montuosa della Palestina, “Joab”, che è stato un antico condottiero del popolo d’israele, “Manassè”, un sovrano, “Jeremia”, un profeta e Sion, il modo in cui chiamano la cosiddetta “terra promessa”. Similmente al Nuovo Testamento, sembra che l’Ebraismo dia molta importanza ai profeti e agli uomini importanti del passato di questa religione, ma di diverso ha che è molto legata al popolo d’Israele, mentre il Cristianesimo è una religione “universale” che si adatta a tutti i popoli.

A questo punto ho analizzato le parole in comune tra i testi sacri.
Effettuando un inner_join sulle parole dei 3 testi sacri, ho ottenuto una tabella contenente soltanto le parole presenti in tutti e 3 i testi, di queste però ho preso soltanto le 10 più frequenti, da cui ho ricavato questo grafico:



La distribuzione di frequenza rappresenta la frequenza all'interno dei singoli libri.

Considerazioni:

Possiamo notare come la parola "God" sia molto presente all'interno della Bibbia e dei Vangeli, mentre di meno nel Corano, in cui "Dio" viene chiamato "Allah".

Diverse parole sono più comuni per Bibbia e Vangeli, come "figlio" e "casa"; la prima stranamente appare di più nell'Antico Testamento, mentre la storia del figlio di Dio viene raccontata effettivamente nel Nuovo Testamento.

Altre parole come "persone", "giorno", "Signore" e "bambini" sono presenti in quantità significative in tutti e 3 i testi.

Da questo risultato, poco interessante, ho deciso di non effettuare una verifica della correlazione su questi 3 testi.

Infine ho analizzato gli n-grams, in particolare ho iniziato con i bigrams, ma ho ottenuto come risultato che tutti i bigram di tutti e 3 i testi appaiono non più di 1 volta all'interno di ciascuno, invece per i trigrams la situazione era diversa, quindi ho deciso di analizzare quelli.

Una volta ricavati i trigrams in comune tra i 3 testi, mi sono reso conto che erano trigrams poco significativi per dedurre informazioni interessanti.

Quelli più significativi, che apparivano più volte all'interno dei 3 testi, erano trigrams del linguaggio comune, nessuno che desse informazioni particolari, quindi ho deciso di non plottarli e di non proseguire ulteriormente con l'analisi degli n-grams.

Per quanto riguarda il primo raggruppamento, non ho ottenuto risultati interessanti e curiosi come nel caso della sentiment analysis; ora effettuerò le stesse operazioni sul secondo raggruppamento che riguarda le religioni orientali.

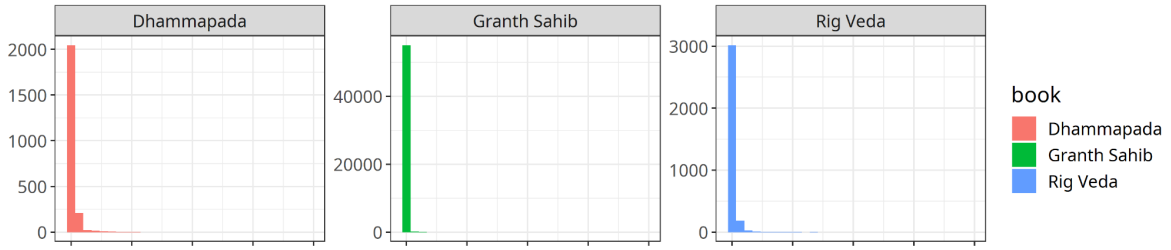
--Religioni Orientali--
Da qui in poi tratto il secondo gruppo, composto da:

- Rig Veda (Induismo)
- Dhammapada (Buddhismo)
- Shri Guru Granth Sahib (anche Adi Granth o Granth Sahib, Sikhismo)

Come al solito ho messo in ordine i dataset ed ho creato una tabella contenente le parole e la loro frequenza assoluta. Alcune stop words le ho rimosse manualmente in quanto, facendo parte di un inglese arcaico (ye, thou, thine, …), non sono presenti nel dataset delle stop words.
La stessa cosa l'ho fatta con diverse parole in lingua originale, ho provato a fare un inner_join con un dizionario inglese ma non sono riuscito a farlo funzionare (probabilmente problemi di metadati).

Dopodichè ho messo insieme tutti i dataset con un full_join, inserendo una colonna per contenere il libro in questione.

Questa prima analisi non considera le parole in comune, ma tutte quante le parole.
Con la tabella ottenuta ho plottato la distribuzione di frequenza delle parole all'interno dei 3 testi:

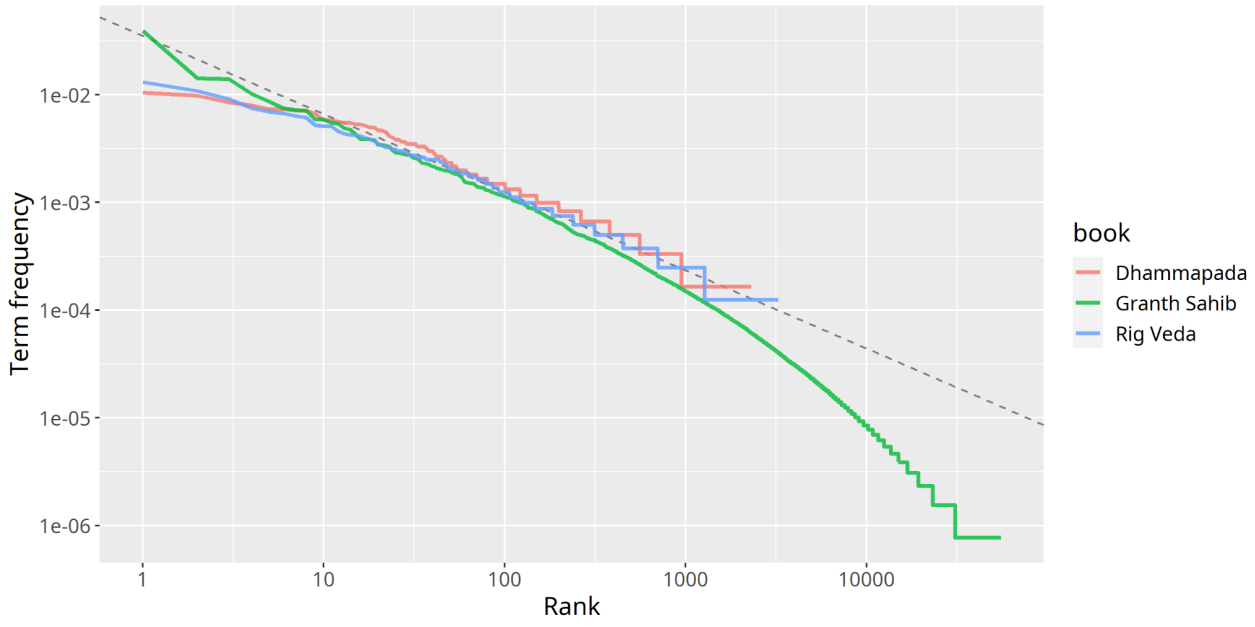


Word distribution among the
3 biggest asian religion's sacred texts

Come si può notare c'è una distribuzione molto comune; poche parole appaiono tante volte e molte appaiono poche volte. Questa relazione è chiamata legge di Zipf.
In questo caso la legge di Zipf è più forte rispetto al caso dei 3 testi precedenti, in quanto le code sono più pesanti e le parole che appaiono poche volte sono molte di più.
Per confermarlo decido di sfruttare questa informazione generando un ranking per le parole:

- Stabilisco un rango per ogni parola in ordine inverso rispetto alla frequenza, lo faccio ordinando il dataset in ordine decrescente rispetto alla frequenza e poi aggiungo una colonna con il numero di riga per il rango.
- Secondo la legge di Zipf, il rango è inversamente proporzionale alla frequenza.

Ottengo una tabella contenente la frequenza di occorrenza delle parole e un ranking associato ad ogni parola, che mi permette di plottare un grafico del rango contro la frequenza:



La scala utilizzata è $\log_{10}$.

Considerazione:

Il grafico ricorda molto una distribuzione classica della legge di Zipf, ci indica quindi che la distribuzione di frequenza è approssimativamente esponenziale.

Poche parole sono ripetute moltissime volte, per tutti e 3 i testi, e molte sono ripetute poche volte.

Le deviazioni ai bassi ranghi indicano che nei testi sacri c'è un uso minore delle parole più comuni del linguaggio, questo probabilmente dovuto all'antichità dei testi.

La curva del Granth Sahib è particolarmente lunga perché il testo è più lungo degli altri 2, quindi contiene più parole e conseguentemente più ranghi.

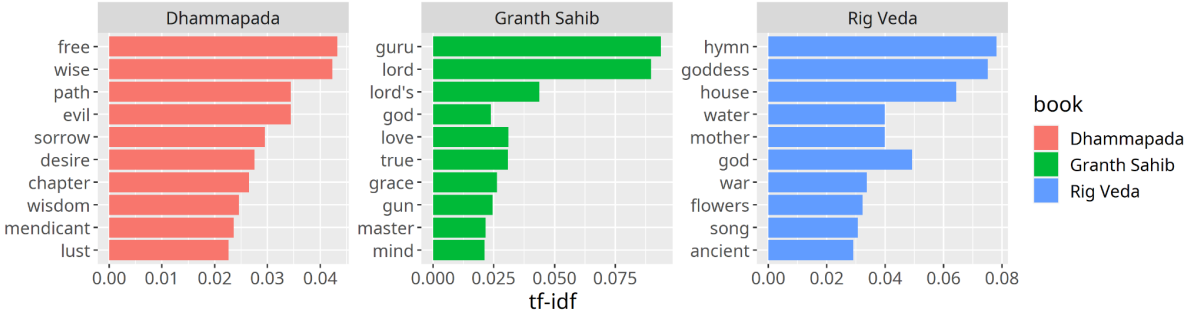
La parte della curva relativa ai bassi ranghi si avvicina di più al modello esponenziale.

Inoltre questa curva assomiglia particolarmente alle curve delle altre 3 religioni monoteiste; questo rappresenta un ulteriore similarità rispetto a quelle riscontrate con la sentiment analysis.

Alla fine di questa sezione effettuerò un'analisi della document term frequency su tutte le religioni monoteiste.

Mi incuriosisce il fatto che dei testi così antichi, rispettino la legge di Zipf come i testi moderni, probabilmente però ciò è dovuto al fatto che questi testi sono stati tradotti, e reinterpretati, più volte nel corso dei secoli.

Ora analizzo la statistica tf-idf dei 3 testi sacri:

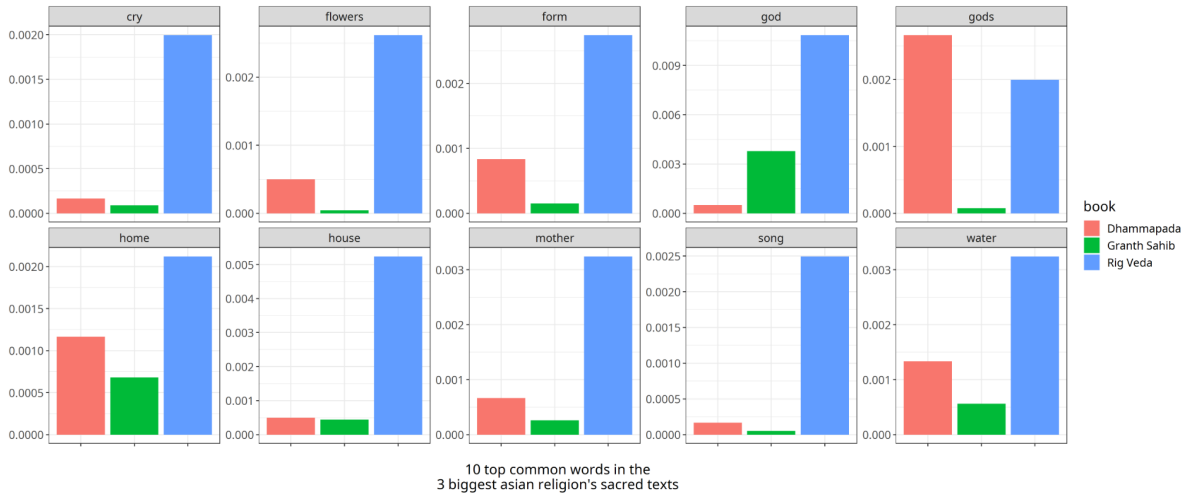


Considerazione:

Non ci sono similarità significative tra i 3 testi, a parte la parola “God”, che per ovvie ragioni è presente in ogni testo di questo tipo.

Procederò comunque con un’analisi delle parole comuni per vedere se scoprirò qualcosa di interessante.

A questo punto analizzo le parole in comune tra i testi sacri.
Effettuando un inner_join sulle parole dei 3 testi sacri, ottengo una tabella contenente soltanto quelle presenti in tutti e 3 i testi, di queste però prendo soltanto le 10 più frequenti in tutti e 3, e ottengo questo grafico:



Considerazione:

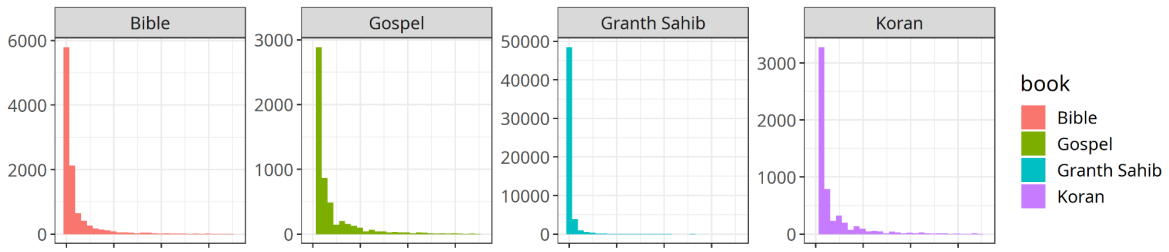
Il Rig Veda ha avuto un'influenza maggiore in questa analisi. Probabilmente lo sbilanciamento presente in questi grafici è dovuto al fatto che le parole in comune sono poche, e le loro distribuzioni all’interno dei singoli libri sono molto diverse. Solamente la parola “casa” presenta una distribuzione simile, leggermente così anche la parola “acqua”, ma non sono particolarmente interessanti dal punto di vista analitico.

Infine ho analizzato gli n-grams, in particolare sono partito con i bigrams, ma ho ottenuto come risultato che tutti i bigram ti tutti e 3 i testi appaiono non più di 1 volta all’interno del testo, invece per i

trigrams la situazione era diversa, quindi ho deciso di analizzare quelli.
Una volta ricavati i trigrams in comune tra i 3 testi, mi sono reso conto che erano trigrams poco significativi per dedurre informazioni interessanti.
Quelli più significativi, che apparivano più volte all'interno dei 3 testi, erano trigrams del linguaggio comune, nessuno che desse informazioni particolari, quindi ho deciso di non plottarli.
L'ultima cosa che ho fatto, è stata quella di cercare attraverso l'analisi degli n-grams quali sono gli dei più menzionati all'interno del Rig Veda, cercando i trigrams contenenti le parole "god of", ma anche qui non ho trovato risultati interessanti, in quanto vengono tutti menzionati non più di una volta.

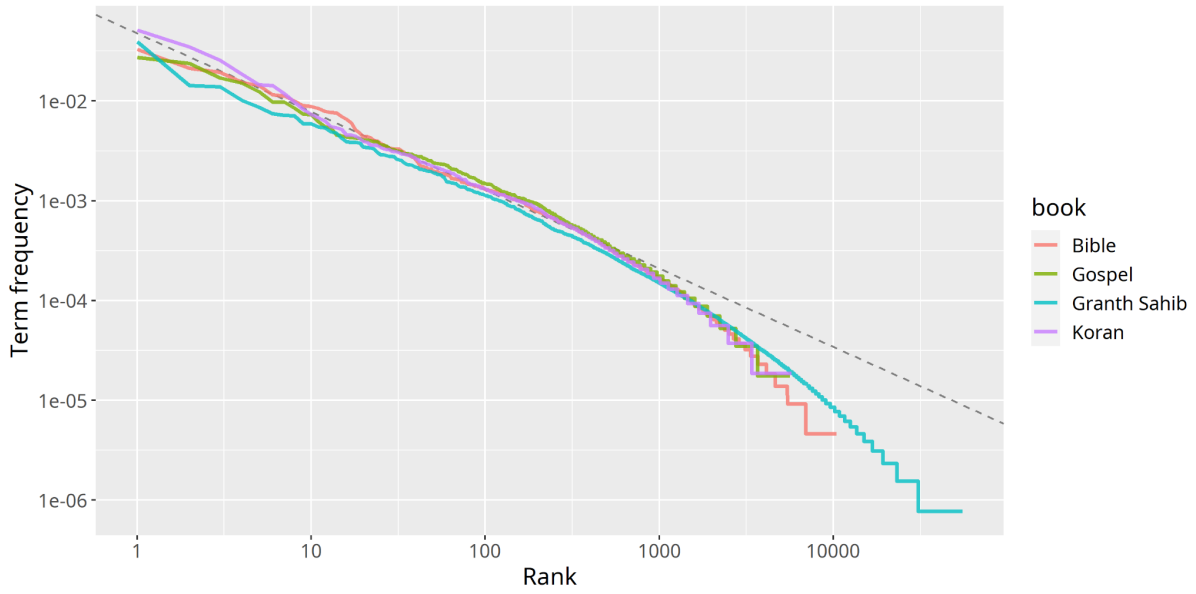
Ora, avendo trovato diverse similarità ed analogie tra le religioni monoteiste, approfondisco l'analisi considerandole in un unico raggruppamento.

Per quanto riguarda la distribuzione di frequenza delle parole all'interno dei 4 testi, questo è il risultato ottenuto:



Word distribution among the
main monotheist religion's sacred texts

Come al solito è presente la legge di Zipf.
Stabiliti i ranghi, questo è il grafico della distribuzione di frequenza contro il rango:



Le distribuzioni di frequenza dei testi sacri delle religioni monoteiste sono molto simili, ma l'analisi tf-idf non ha portato alcun risultato interessante, così come l'analisi delle parole comuni.
Apparentemente le 3 grandi religioni monoteiste sono particolarmente simili sia per quanto riguarda il sentiment che per quanto riguarda la word and document term frequency, ma il Granth Sahib di somigliante ha solamente la distribuzione di frequenza delle parole ed il sentiment, ma non ci sono parole comuni particolarmente interessanti.

Conclusioni:

- Nonostante l'antichità dei testi, questi ultimi hanno comunque una distribuzione di frequenza delle parole molto simile ai testi odierni e rispettano anche la legge di Zipf, ma questo è probabilmente dovuto al fatto che sono stati tradotti e reinterpretati più volte nel corso dei secoli.
- I testi sacri delle 3 grandi religioni monoteiste presentano poche similarità, mentre i testi sacri delle 3 più grandi religioni asiatiche nessuna.

È interessante il fatto che l'antico Testamento presenta moltissimi riferimenti alla storia del popolo Ebraico, ciò indica che la religione è fortemente legata a questo popolo, mentre i riferimenti più presenti all'interno del Nuovo Testamento sono più generici, questo fa sì che la religione Cristiana si adatti bene a qualsiasi etnia o popolo.

- Il Corano presenta la parola "Allah", quindi "Dio", molto più frequentemente di Antico e Nuovo Testamento, che invece danno ampio spazio a personaggi come profeti e re, nel caso dell'Antico Testamento, apostoli e familiari di Gesù Cristo nel caso del Nuovo Testamento.
- Il Granth Sahib presenta una distribuzione di frequenza molto più simile ai testi sacri delle 3 grandi religioni monoteiste rispetto ai testi sacri delle più grandi religioni asiatiche, ma per quanto riguarda le parole comuni non c'è nessun dato interessante.
- Speravo di poter osservare quali sono gli dei più menzionati nel Rig Veda, ma ho trovato che vengono tutti menzionati una sola volta. La curiosità in questo sta sempre nel fatto che la "ricerca dell'equilibrio interiore" è uno dei principi fondamentali di queste 2 religioni e questo equilibrio nel menzionare tutti gli dei allo stesso modo rappresenta un risultato interessante.

Risposte:

1. Quali sono le parole più ricorrenti all'interno dei testi sacri?
   La parola "Dio" ovviamente è quella che ricorre di più in tutte le religioni, però un dato interessante è che l'Antico Testamento presenta molti riferimenti a personaggi importanti nella storia del popolo Ebraico, e questo lega fortemente la religione al popolo. Mentre per il nuovo Testamento e per le altre religioni i riferimenti sono più generici, facendo quindi sì che le altre religioni si adattino bene a qualsiasi popolo o etnia.

Il corano presenta la parola “Allah” quindi “Dio” più volte rispetto ad Antico e Nuovo Testamento, che danno più spazio invece a profeti e personaggi.

2. Testi sacri di religioni molto differenti, hanno parole in comune?
   No, l’unica relazione che c’è tra i testi sacri di religioni molto differenti è quella che c’è tra i testi delle 3 grandi religioni monoteiste ed il Granth Sahib, che hanno distribuzioni di frequenza molto simili, ma nessuna parola in comune particolarmente rilevante. La curiosità è che si parla di 4 religioni monoteiste.
3. In caso ci siano, può essere ciò dovuto ad una qualche origine in comune tra le religioni? Per esempio la posizione geografica?
   Ci sono similarità tra le 3 grandi religioni monoteiste, che sono legate da radici comuni e zone geografiche di origine molto vicine, mentre per quanto riguarda la relazione che lega queste religioni ed il Sikhismo non è forte, ed inoltre si tratta di 2 zone geografiche e culturali completamente differenti. Direi che questa relazione non c’è, ma è curioso il fatto che sono 4 religioni monoteiste, e qualcosa in comune esiste, mentre rispetto a Buddhismo e Induismo sono completamente diverse.

3. Topic Modeling

In quest'ultima parte voglio analizzare gli argomenti trattati e voglio trovare delle somiglianze tra gli argomenti dei 6 testi sacri che ho trattato, raggruppati nei seguenti modi:

- 3 testi sacri delle grandi religioni monoteiste (Bibbia, Vangelo, Corano)
- 2 testi sacri delle religioni nate in palestina (Bibbia, Vangelo)
- 3 testi sacri delle più grandi religioni orientali (Rig Veda, Dhammapada, Granth Sahib)
- 2 testi sacri delle religioni Induismo e Buddhismo (Rig Veda, Dhammapada)
- 4 testi sacri delle 4 religioni monoteiste

Domande:

1. Quali sono gli argomenti più presenti?
2. Ci sono argomenti in comune tra i testi sacri di religioni molto differenti?
3. In caso positivo, questi risultati possono rafforzare le ipotesi fatte in precedenza (sono risultati significativi)?

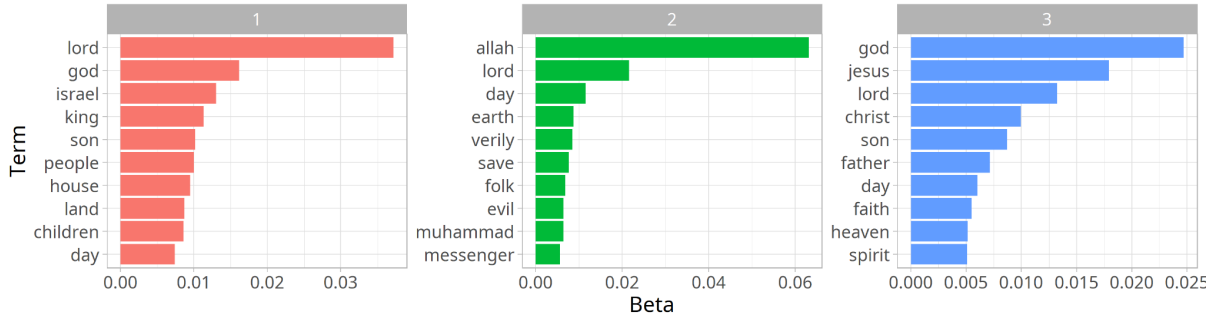
Analisi:

--3 Grandi Religioni Monoteiste--

Per prima cosa ho messo in ordine i vari dataset, alcune parole le ho rimosse manualmente poichè erano presenti parole dell'inglese antico come "ye", "thou", "thy", ecc… che non erano presenti nel dataset delle stop words.

Dopodichè ho unito i 3 libri in un unico dataset, contraddistinti da un valore "document", e ho creato una document-term matrix.

A questo punto ho i dati nel formato che mi interessa e posso procedere con la Latent Dirichlet Allocation, che mi permette di calcolare la word-topic probability e di plottare i seguenti grafici:



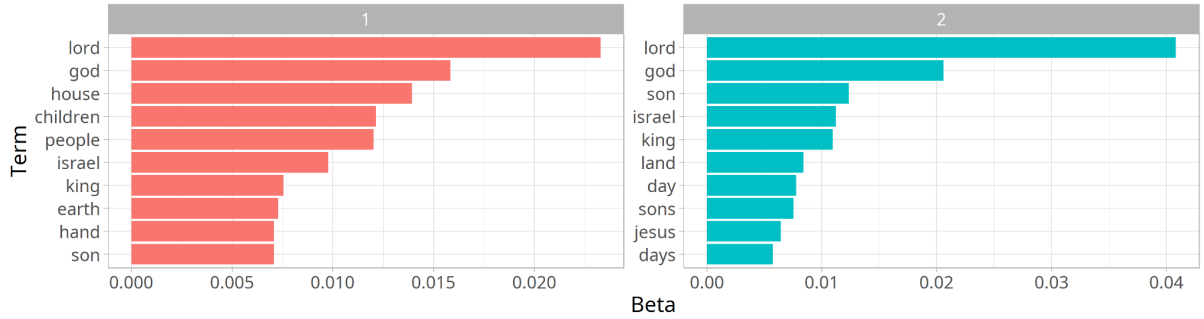
Qui ho deciso di estrapolare 3 topic, principalmente perchè voglio vedere se sono in grado di distinguere i 3 topic nei 3 libri, e trovare similarità o differenze.
Il grafico contiene le 10 parole più frequenti in ciascun topic.

Considerazione:

L'algoritmo ha trovato 3 topic ed è molto probabile che i 3 topic principali siano distinguibili proprio nei 3 testi sacri.
Il 2 è sicuramente relativo al Corano per la presenza delle parole "Allah" e "Maometto", il 3 ai Vangeli per la presenza della parola "Gesù" e "Cristo", e il numero 1 quindi ad esclusione sarà relativo alla Bibbia.
Inoltre è presente la parola "Land" che rimanda a tutta la questione della ricerca della "terra promessa" narrata nell'Antico Testamento.
La parola "Dio" ricorre in tutti e 3 i topic, anche se nel Corano sotto forma di "Allah". Un'altra parola, "Signore", è caratteristica di tutti e 3 i topic.
Anche la parola "giorno" lega in qualche modo tutti e 3 i topic.
Una stranezza è che la parola "figlio", presente nel primo e terzo topic (che ho riconosciuto riguardare rispettivamente Bibbia e Vangeli), ha una probabilità per-topic-per-word più alta nel

primo topic rispetto al terzo topic (> 0.01 vs < 0.01), quando la storia del “figlio di Dio” viene raccontata nel Vangelo.

Ora creo una document-term matrix contenente solamente le parole della Bibbia e del Vangelo, e proseguo con la stessa identica procedura usata in precedenza ed infine ottengo i seguenti grafici:



Per lo stesso motivo precedente ho considerato 2 topic e credo che sia abbastanza evidente che il topic 1 riguarda principalmente la Bibbia ed il 2 il Vangelo, dato che nel 2 è presente la parola “Gesù”.

Considerazione:

Le parole “Signore” e “Dio” hanno il beta più alto in entrambi i topic.

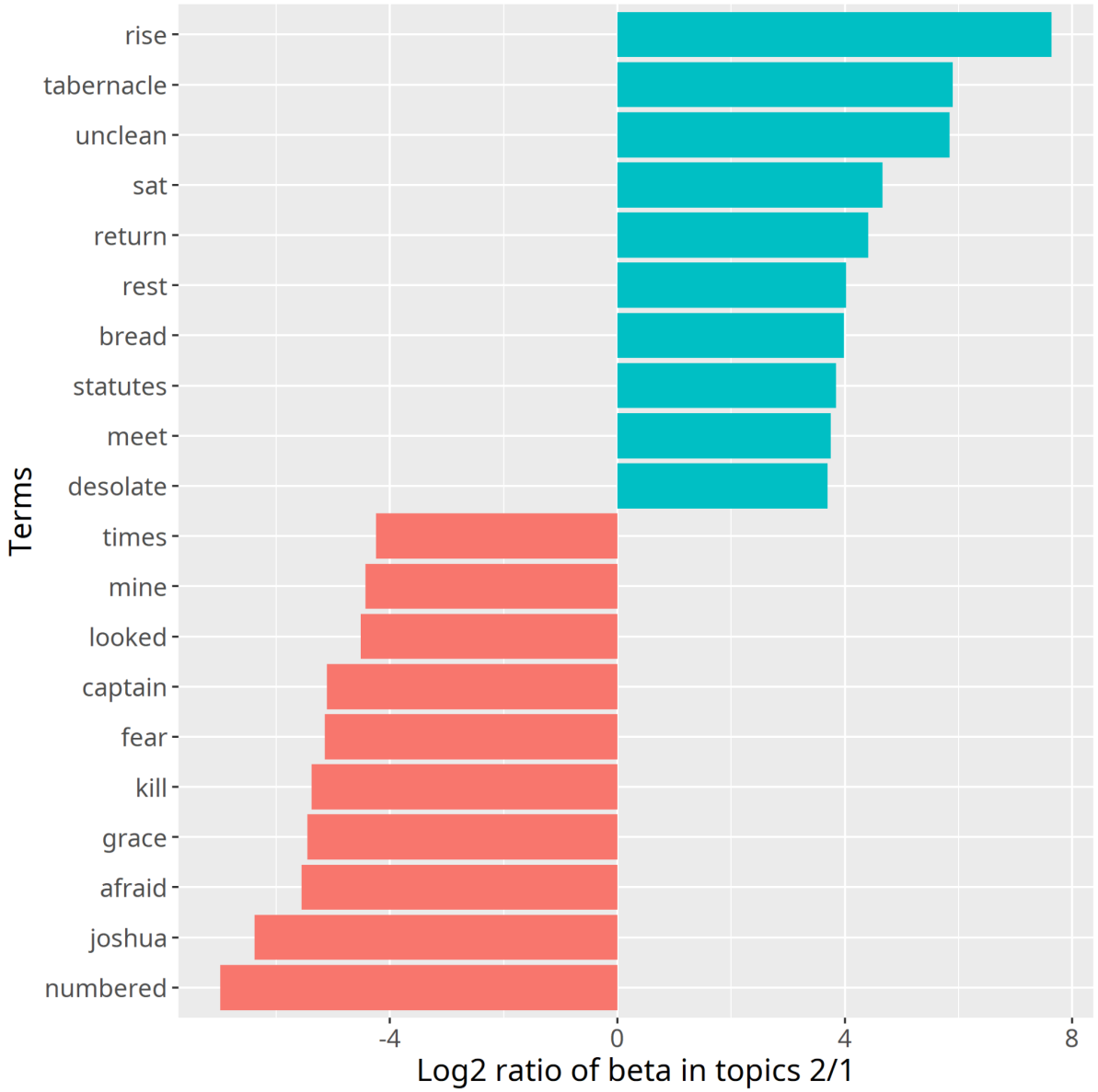
Diversamente da prima, la parola “figlio” ha un beta più alto nel topic che io ho associato al Nuovo Testamento. Probabilmente la presenza del Corano ha influenzato l’analisi precedente.

La parola “figlio” è probabilmente più presente all’interno del Nuovo Testamento come si pensa che sia.

La parola “Israele” è presente in entrambi i topic, ma ha un beta leggermente più alto nel secondo, ma si potrebbe comunque dire che è abbastanza equilibrato.

Da queste analisi evinco che la somiglianza tra Antico e Nuovo Testamento esiste, è abbastanza forte ed è sicuramente legata al fatto che il Cristianesimo derivi dall’Ebraismo.

Ora, invece plotto il rateo logaritmico delle per-term-per-document probabilities, per analizzare meglio le differenze:



Considerazione:

Parole come "ascesa", "tabernacolo", "pane", "ritorno" sono più caratteristiche del topic 2, e sono parole che si associano perfettamente al Nuovo Testamento, mentre parole come "Giosuè", un condottiero Ebraico, "uccidere", "paura", "grazia" appartengono di più al topic 1.

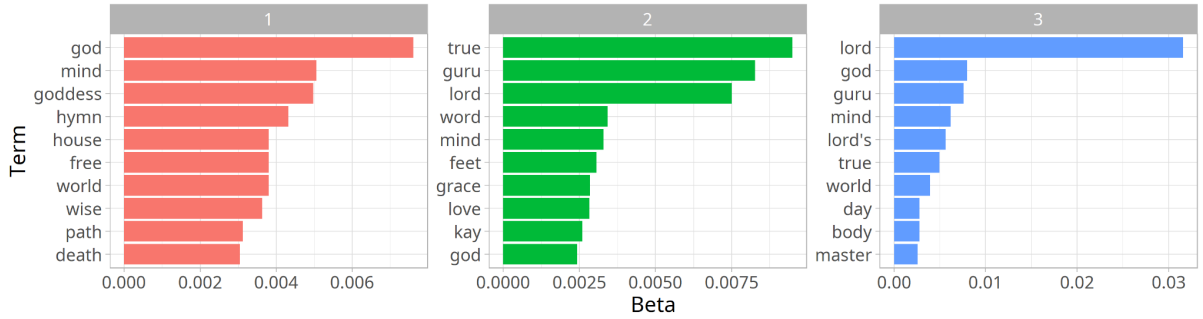
Questo risultato mi porta a pensare ancora di più che il topic 1 sia associato alla Bibbia ed il 2 ai Vangeli.

--Religioni orientali--

Per prima cosa ho messo in ordine i vari dataset, alcune parole le ho rimosse manualmente poichè erano presenti parole dell'inglese antico come "ye", "thou", "thy", ecc… che non erano presenti nel dataset delle stop words.

Dopodichè ho unito i 3 libri in un unico dataset, contraddistinti da un valore "document", ed ho creato una document-term matrix.

A questo punto ho i dati nel formato che mi interessa e posso procedere con la Latent Dirichlet Allocation, che mi permette di calcolare la word-topic probability e di plottare i seguenti grafici:



Qui ho deciso di estrapolare 3 topic, principalmente perchè voglio vedere se sono in grado di distinguere i 3 topic nei 3 libri, e trovare similarità o differenze.

Il grafico contiene le 10 parole più frequenti in ciascun topic.

Considerazione:

L'algoritmo ha rilevato i 3 topic, che anche in questo caso sono assimilabili ai 3 testi sacri, anche se il riconoscimento non è così evidente come nel caso precedente. Non conoscendo queste religioni (nel caso precedente si poteva distinguere il nuovo Testamento dalla parola "Gesù" ed il Corano dalla parola "Allah"), qui abbiamo 2 testi che sono molto simili tra di loro, Dhammapada e Rig Veda. La parola Buddha non è presente nella top 10 delle parole che contraddistinguono i topic, quindi non saprei dire quale topic riguarda il Dhammapada, ma la parola "kay" riguarda sicuramente il Granth Sahib, in particolare

si riferisce alle “cinque K”, che sono alcuni dei principi fondamentali di questa religione:

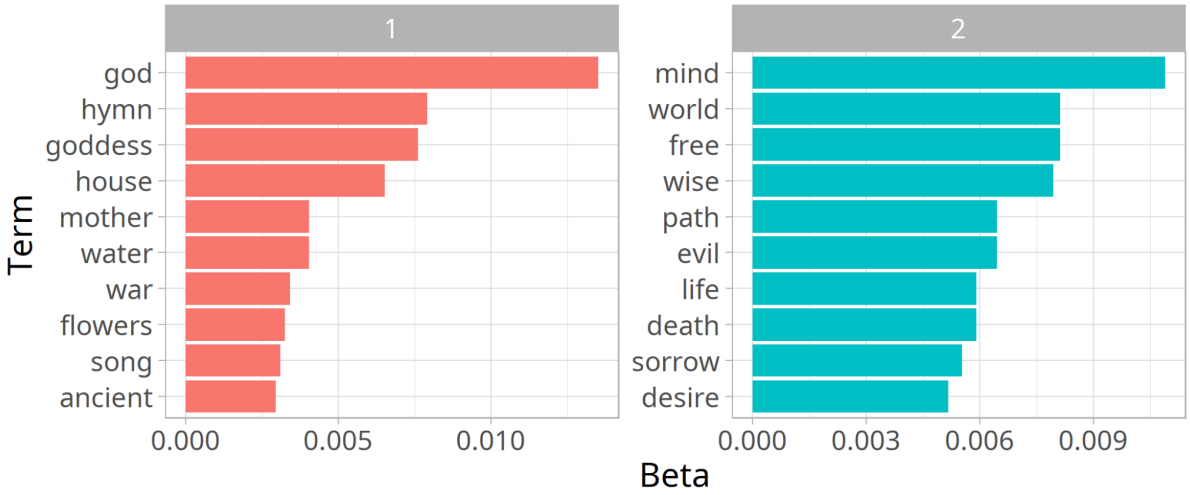
- Kesh: divieto di tagliare i capelli, la barba e i peli del corpo (i primi sono coperti dal “dastar”, un particolare turbante). Queste parti del corpo rappresentano rispettivamente la virilità, il coraggio e la saggezza.
- Kangha: piccolo pettine di legno che serve a tenere i capelli fermi sotto il turbante, simbolo di pulizia.
- Kachera: pantaloni da combattimento, larghi, utili per andare a cavallo senza disturbare il movimento.
- Kara: un braccialetto di metallo simboleggiante l’umiltà e l’appartenenza al divino.
- Kirpan: Un pugnale, vero o in miniatura, simbolo di coraggio. Serve al sikh per tenere a mente le persecuzioni per cui la loro religione e molte altre hanno sofferto e la necessità di difendere la libertà di coscienza (la loro e quella degli altri) contro l'oscurantismo.

Anche la parola “grazia” è caratteristica solamente di questo testo, quindi deduco che il topic 2 riguarda principalmente il Granth Sahib.

Per quanto riguarda gli altri 2 topic abbiamo la parola “saggio”, che è presente solamente all’interno del Dhammapada, quindi deduco si tratti del topic 1, ma la parola “goddess” non è presente, e stessa cosa per la parola “guru” del topic 3, quindi questi due topic non sono ben identificati.

Questa unione, che potrebbe essere stata creata nei topic tra Dhammapada e Rig Veda, può essere stata causata da una sostanziale somiglianza tra i 2 testi, che sono anche relativi alle 2 religioni Induismo e Buddhismo, la seconda derivata dalla prima, anche se la prima è una religione politeista, e la seconda è una dottrina filosofica che non venera dei ma si concentra sull’essere umano.

Ora analizzo solamente Dhammapada e Rig Veda, come già fatto prima ed ottengo questi grafici:



Considerazione:

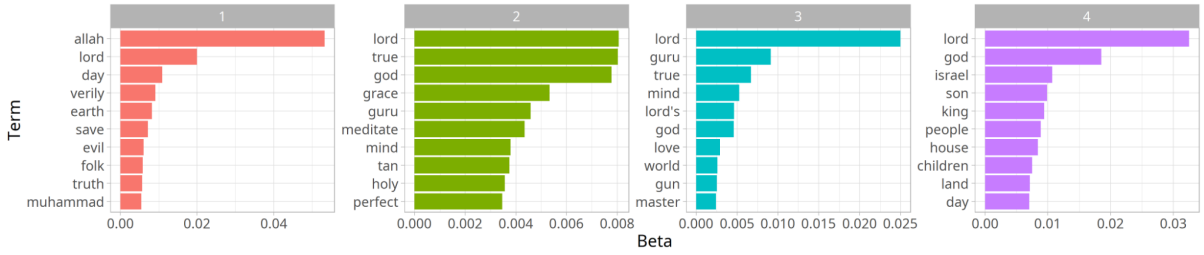
In questo caso abbiamo la parola “guerra” che è caratteristica del Rig Veda, e le parole “saggio” e “tristezza” che sono caratteristiche del Dhammapada, quindi i due topic probabilmente si riferiscono principalmente a Rig veda per il topic 1 e Dhammapada per il topic 2.

Dalla mia analisi precedente non sono così sicuro che i due topic si riferiscano ai due libri distintamente; questi ultimi hanno diverse cose in comune, ma le due religioni, seppur una derivata dall’altra, sono sostanzialmente differenti nei credi e nei concetti.

Non ho considerato il grafico del rateo logaritmico delle per-term-per-document probabilities, per analizzare meglio le differenze, poichè le parole presenti erano poco interessanti per la mia analisi.

Questa cosa rinforza la mia tesi sul fatto che Dhammapada e Rig Veda si assomiglino, e così le due religioni Buddhismo ed Induismo.

Infine ho considerato le 4 religioni monoteiste ed i loro testi sacri, la per-word-per-topic probability distribution è la seguente:



Considerazione:

Come al solito considero un topic per libro, il primo topic è sicuramente influenzato maggiormente dal Corano, mentre per gli altri 3 la situazione cambia.

Il terzo ed il quarto topic hanno distribuzioni molto simili ai grafici visualizzati precedentemente per Granth Sahib ed Antico Testamento, in più nel quarto sono presenti “Israele” e “Terra” che caratterizzano, come abbiamo visto, la Bibbia. Il topic 2 quindi mi viene da pensare che sia stato principalmente influenzato dal Nuovo Testamento, ma appare la parola “guru” che è fuorviante.

Evidentemente il nuovo Testamento assimila topics anche presenti negli altri libri, a rinforzare la tesi secondo la quale il “Cristianesimo” è una religione che si adatta bene a diversi popoli, diverse etnie e diverse culture, mentre le altre religioni monoteiste sono separate da barriere culturali e storiche evidenti.

Conclusioni:

- Da quest’ultima analisi si evince una cosa in particolare, ovvero che le religioni che sono “una il derivato dell’altra”, oppure dei “distaccamenti” dalle religioni “originali”, portano nei testi sacri diverse somiglianze, argomenti comuni, ma dimostrano anche di avere delle differenze:
    - La Bibbia ed i Vangeli provengono entrambe dalla religione Ebraica, da cui poi è nato il Cristianesimo. Gli argomenti trattati nei testi presentano molte similarità

grazie a questa radice comune, ma sono visibili anche le differenze, come il fatto che la prima è fortemente legata alla storia del popolo Ebraico, e la seconda è fortemente legata alla storia di Gesù Cristo.

- Dhammapada e Rig Veda provengono entrambe da una radice Induista, da cui il Buddhismo si è distaccato professando una filosofia antropocentrista ed andando contro al concetto di "divinità". I due testi presentano diversi argomenti in comune ma sostanziali differenze.

- La seconda tesi che ho rafforzato con questa analisi è che il Nuovo Testamento contiene topic adattabili bene anche agli altri testi delle religioni monoteiste, e ciò rinforza la mia tesi precedente secondo la quale il Cristianesimo è una religione che si può adattare bene a qualsiasi popolo ed etnia, non per altro è la religione più diffusa al mondo (con le sue varianti), ed è professata da popoli ed etnie di culture agli antipodi.

Risposte:

1. Quali sono gli argomenti più presenti?
   Sicuramente "Dio" è l'argomento principale, ma c'era da aspettarselo, per i Vangeli, la storia di Gesù Cristo e dei suoi seguaci (apostoli), mentre per quanto riguarda la Bibbia, la storia del popolo di Israele, le dodici tribù, e le varie guerre che hanno dovuto affrontare per ottenere "la terra promessa".
   Per quanto riguarda le altre 3 religioni, anche qui "Dio" è sostanzialmente l'argomento più presente, anche nel Dhammapada che professa una filosofia antropocentrica, che si distacca dal concetto di "divinità", il Granth Sahib contiene diverse argomentazioni riguardanti le armi, la guerra, ed il combattimento, (questo l'ho potuto comprendere meglio anche dalla mia piccola ricerca sulle 5 K), infine il Rig Veda contiene principalmente argomenti riguardanti i vari dei della religione

politeista, e diversi riferimenti alla natura, probabilmente legata agli stessi.

2. Ci sono argomenti in comune tra i testi sacri di religioni molto differenti?
   Si, in particolare per il Nuovo Testamento, che presenta similarità con le altre religioni, rendendo la religione Cristiana potenzialmente adattabile ad etnie, popoli e culture diversi.
3. In caso positivo, questi risultati possono rafforzare le ipotesi fatte in precedenza (sono risultati significativi)?
   Si.

Sitografia:

Datasets:

- Bibbia.txt
- Corano.txt
- Rig Veda.txt
- Dhammapada.txt
    - Granth Sahib.pdf

Religioni per nazione - Wikipedia
Religioni maggiori - Wikipedia (da qui si raggiungono tutte le altre pagine che ho utilizzato)

Documentazione R:

- R Markdown: The Definitive Guide (bookdown.org)
- The Comprehensive R Archive Network (r-project.org)

Della Rovere Sandro Junior
147687