README.md

# Relazione del secondo gruppo di esercizi

## TLN_dicaro_2.1

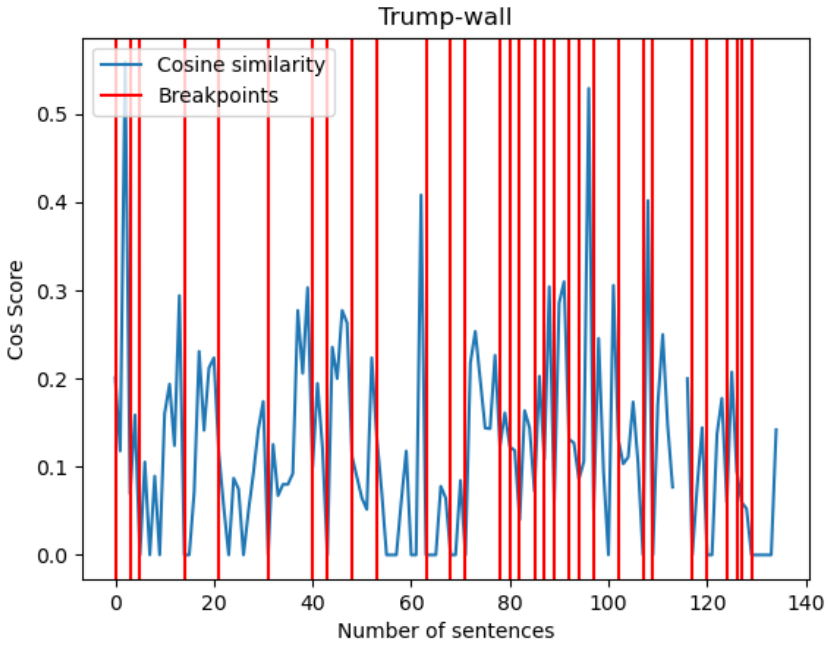
*CONSEGNA*:

- Ispirandosi al text-tiling, implementare un algoritmo di segmentazione del testo.
- Sfruttare informazioni come le frequenze e le co-occorrenze ed eventuale pre-processing del testo.

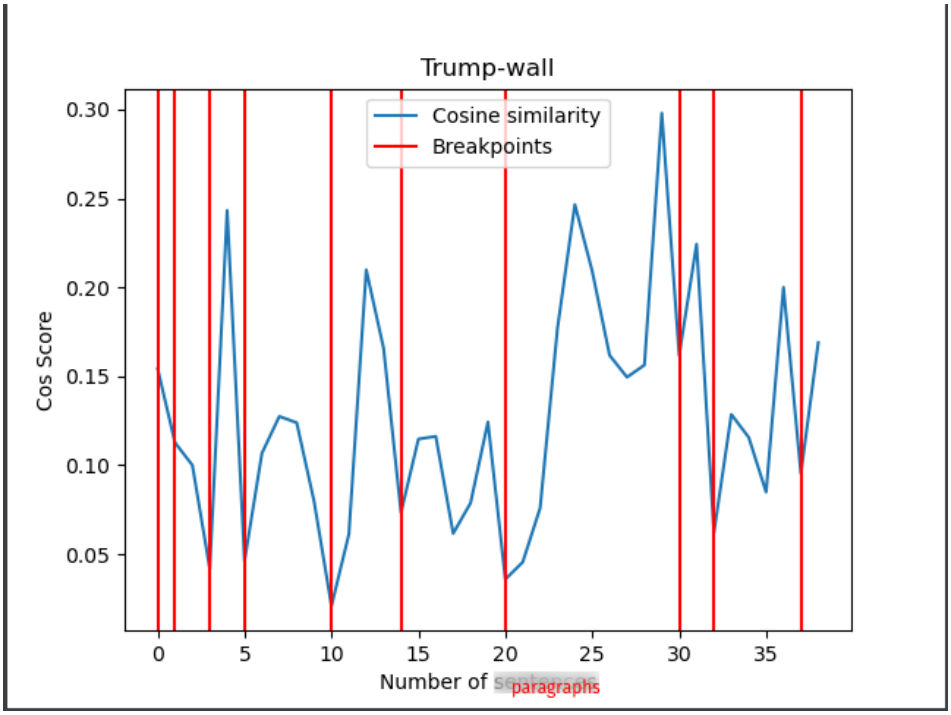
*SVOLGIMENTO*:

- Breve lettura iniziale del seguente articolo
- Si è scelto di utilizzare un testo su Trump-wall già usato nella seconda parte del corso come file d' ingresso
- Si crea per prima cosa il dizionario del testo, annotando le frequenze di ogni parola (previo filtering e lemmatizzazione).
- Si creano poi i vettori che contengono i termini presenti in ogni frase.
- Si calcola la cosine similarity fra tutti questi vettori colonna.
- Si itera su queste cosine similarities per identificare i punti in cui il valore scende al di sotto della loro media di una certa percentuale: in questi punti viene inserito un punto di cambio di discorso.
- L' operazione viene ripetuta per un certo numero d' iterazioni, usando come confronto la media di quel preciso segmento invece che la media complessiva.

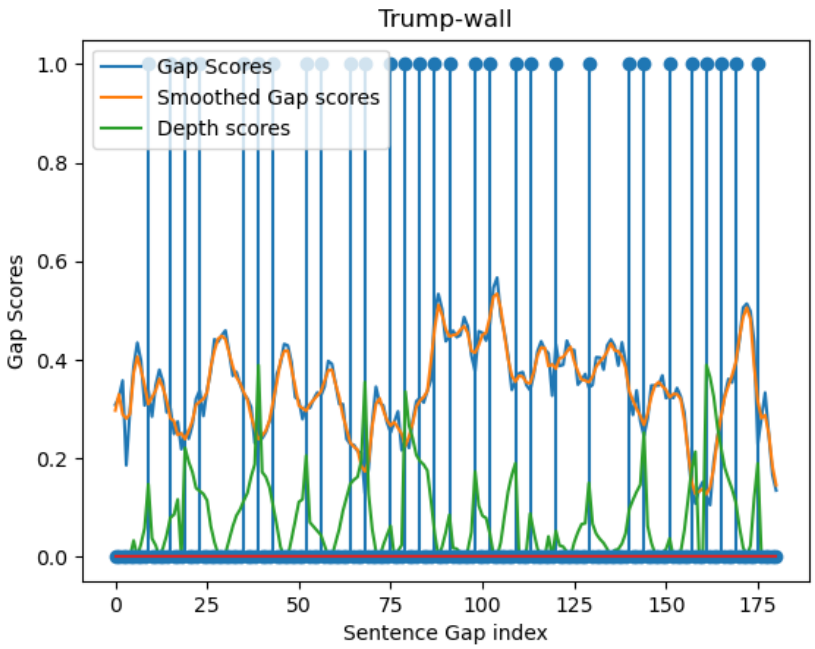
*RISULTATI*:




- Notiamo un valore della cos sim molto altalenante, anche a causa di frasi dalla lunghezza altamente variabile




- Il discorso migliora se anziché lavorare sulle singole frasi, passiamo ai paragrafi.
- Questo approccio è però meno generalizzabile poi dipende molto dalla formattazione del testo originale




Come confronto ulteriore si usa implementazione di nltk del TextTiling disponibile qui

- I due grafici non sono del tutto sovrapponibili per via di come gestiscono il file ingresso (in questo caso in maniera *raw*)

### TLN_dicaro_2.2

*CONSEGNA*:

- Topic modeling partendo da un corpus con visualizzazione

*SVOLGIMENTO e RISULTATI*:

- Come corpus abbiamo utilizzato il 20-Newsgroup dataset caricato come file .json. Questa versione contiene circa 11.000 newsgroups appartenenti a 20 diversi topics. Lo abbiamo importato come data frame la cui testa appare in questo modo:

| | content | target | target_names |
|---|---|---|---|
| 0 | From: lerxst@wam.umd.edu (where's my thing)\nS... | 7 | rec.autos |
| 1 | From: guykuo@carson.u.washington.edu (Guy Kuo)... | 4 | comp.sys.mac.hardware |
| 2 | From: twillis@ec.ecn.purdue.edu (Thomas E Will... | 4 | comp.sys.mac.hardware |
| 3 | From: jgreen@amber (Joe Green)\nSubject: Re: W... | 1 | comp.graphics |
| 4 | From: jcm@head-cfa.harvard.edu (Jonathan McDow... | 14 | sci.space |

- Abbiamo estratto i content e li abbiamo puliti, rimuovendo caratteri inutili in modo da renderli ottimali per i metodi di pre-processing della libreria gensim.
- Tramite gensim (in particolare gensim.models.phrases) abbiamo generato bi-grammi e tri-grammi puliti da stop words.
- Abbiamo fatto la lemmatizzazione del testo considerando solo nomi, aggettivi, verbi e avverbi e, sempre grazie a librerie di gensim, abbiamo associato un id univoco a ogni parola del documento; abbiamo cioè creato un dizionario e poi un corpus per mappare id e frequenza specifica.

**Latent Dirichlet Allocation**

- Ottenuti dizionario e corpus, sempre usando gensim, abbiamo applicato il modello LDA ottenendo la suddivisione in topics. Ogni topic risulta essere una combinazione di keywords che contribuiscono con un certo peso.
- Abbiamo cercato di dare un'interpretazione di senso ai topics usando **WordNet**;
- in particolare abbiamo ottenuto le otto parole più rappresentative per ogni topic e ne abbiamo estratto iponimi e iperonimi e per ogni topic abbiamo estratto il synset più significativo sulla base delle occorrenze tra iponimi e iperonimi comuni.

I risultati *non* sono stati però molto incoraggianti, rivelando una probabile incoerenza interna ai topic.

- Abbiamo calcolato perplexity (*quanto bene il modello rappresenta o riproduce le statistiche dei dati forniti*) e coherence (*grado di similarità semantica tra le parole con alto score nel topic*) per dare una valutazione del modello

| metric | score |
|---|---|
| Perplexity | -8.57 |
| Coherence | 0.41 |

**La bassa coerenza all'interno dei topics significa che probabilmente 20 topics sono troppi e sarebbe necessaria un'ottimizzazione.**

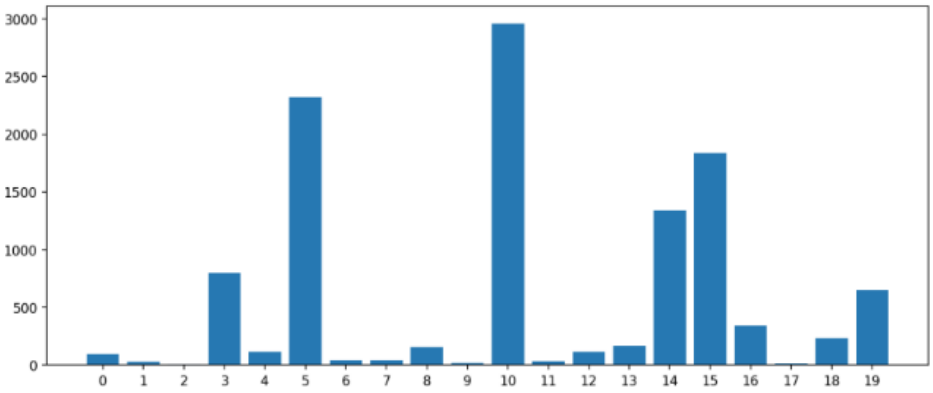
- In questa prospettiva abbiamo, dunque, estratto i cinque lemmi più rappresentativi di ogni content; per la prima cinquina di content questo è il risultato:

| | Topic_Num | Topic_Perc_Contrib | Keywords | Text |
|---|---|---|---|---|
| 0 | 0.0 | 0.9810 | church, subject, would, line, question, write,... | [subject, migraine, organization, univ, line, ... |
| 1 | 1.0 | 0.9220 | entry, program, line, file, rule, size, use, s... | [subject, reposte, international_obfuscated, r... |
| 2 | 2.0 | 0.5265 | color, monitor, line, compression, imake, vram... | [line, wonderful, sale, also, include, disk, d... |
| 3 | 3.0 | 0.9939 | say, people, believe, write, know, subject, th... | [line, respond, moderator, write, choose, beli... |
| 4 | 4.0 | 0.9956 | space, bike, orbit, mission, satellite, line, ... | [wizzard, old, audio_visual, equipment, nanaim... |

- Ogni documento è composto da più argomenti. Ma in genere solo uno degli argomenti è dominante.
- Abbiamo estratto questo argomento dominante per ogni content con il relativo peso e keywords; per il primi 10 contents questo è il risultato:

| | Document_No | Dominant_Topic | Topic_Perc_Contrib | Keywords | Text |
|---|---|---|---|---|---|
| 0 | 0 | 15.0 | 0.9275 | line, car, would, write, subject, use, article... | [where, thing, subject, car, nntp_poste, host,... |
| 1 | 1 | 10.0 | 0.5168 | use, line, subject, file, window, system, writ... | [subject, clock, poll, final, summary, final, ... |
| 2 | 2 | 10.0 | 0.3573 | use, line, subject, file, window, system, writ... | [subject, question, engineering, computer, net... |
| 3 | 3 | 14.0 | 0.7819 | line, write, team, game, subject, year, articl... | [division, line, host, amber, write, write, ar... |
| 4 | 4 | 10.0 | 0.3563 | use, line, subject, file, window, system, writ... | [question, organization, smithsonian_astrophys... |
| 5 | 5 | 5.0 | 0.9855 | would, write, say, people, go, think, line, kn... | [foxvog_dougla, subject, reword, vtt, line, ar... |
| 6 | 6 | 10.0 | 0.7275 | use, line, subject, file, window, system, writ... | [brain, tumor, treatment, thank, people, respo... |
| 7 | 7 | 18.0 | 0.8981 | drive, scsi, chip, line, go, wire, bit, get, s... | [subject, scsi, organization, line, nntp_poste... |
| 8 | 8 | 10.0 | 0.8147 | use, line, subject, file, window, system, writ... | [subject, win, icon, help, line, win, download... |
| 9 | 9 | 15.0 | 0.4890 | line, car, would, write, subject, use, article... | [subject, sigma_design, double, article, write... |

- Abbiamo poi ottenuto un grafico con la distribuzione dei topics all'interno del documento




- Abbiamo poi ottenuto la rappresentazione dei lemmi più significativi per ciascuno dei 20 topic; il risultato per i primi 10 topics è il seguente:

| | Topic_Num | Topic_Perc_Contrib | Keywords | Representative Text |
|---|---|---|---|---|
| 0 | 0.0 | 0.9810 | church, subject, would, line, question, write, make, say, new, existence | [subject, migraine, organization, univ, line, article, write, ask, alternative, try, state, subl... |
| 1 | 1.0 | 0.9220 | entry, program, line, file, rule, size, use, section, build, must | [subject, reposte, international_obfuscated, rule, gmt, reply, line, receive, number, request, r... |
| 2 | 2.0 | 0.5263 | color, monitor, line, compression, imake, vram, would, subject, run, vegetarian | [line, wonderful, sale, also, include, disk, drive, color, moniter, great, shape, software, joys... |
| 3 | 3.0 | 0.9939 | say, people, believe, write, know, subject, think, line, would, thing | [line, respond, moderator, write, choose, believe, rely, important, area, personal, sovereignty,... |
| 4 | 4.0 | 0.9956 | space, bike, orbit, mission, satellite, line, motorcycle, ride, subject, rider | [wizzard, old, audio_visual, equipment, nanaimo, campus, subject, correction, last, followup, li... |
| 5 | 5.0 | 0.9962 | would, write, say, people, go, think, line, know, subject, article | [subject, yet, rushdie, islamic_law, nntp_poste, host, organization, write, understanding, gener... |
| 6 | 6.0 | 0.9923 | msg, food, use, eat, system, taste, line, marriage, superstition, formula | [subject, superstition, line, write, write, add, fuel, flame, war, read, natural, source, mentio... |
| 7 | 7.0 | 0.9360 | gun, safety, wiring, ground, outlet, glock, revolver, gfci, advertising, publish | [subject, need, advice, doctor, patient, relationship, problem, nntp_poste, line, sound, heart, ... |
| 8 | 8.0 | 0.9880 | say, turkish, people, armenian, work, turk, year, child, government, know | [serdar_argic, subject, consider, reply, line, article, write, letter, chronicle, date, figment,... |
| 9 | 9.0 | 0.9441 | trade, captain, line, go, flyer, knife, motto, season, subject, checker | [institute, line, host, fledgling, originator, hate, seat, instead, logistician, go, tiger, go, ... |
| 10 | 10.0 | 0.9940 | use, line, subject, file, window, system, write, program, thank, problem | [access, distribution_usa, line, write, be, study, follow, type, user, would, like, manager, bas... |

- Abbiamo infine ottenuto una word could per rappresentare le parole più rappresentative di ogni topic in diversi colori e grandezza a seconda della rilevanza:


Topic 0
make would
existence church
write
say subject
new
line question
Topic 1
size
rule
program use build
entry must line
file section
Topic 2
subject color run
imake
vram line
compression
vegetarian monitor
would
Topic 3
say
would subject
write believe
think
thing line
people
know
Topic 4
motorcycle
rider orbit
ride
line subject
bike satellite
space
mission
Topic 5
line say go
people know
would
article subject
write think
Topic 6
food taste
superstition
msg eat
line formula use
system
marriage
Topic 7
ground
outlet gun
revolver
publish wiring
gfci glock
advertising safety
Topic 8
armenian
say people
turkish
year
turk know work
child government
Topic 9
subject season
motto knife
flyer captain
go
trade line
checker
Topic 10
problem
file window
line
write
subject
system
thank
program use
Topic 11
make comet
organization gehrel
subject
spline write
comp_graphic
line
balloon
Topic 12
write
subject
would line
mouse saturn
price buy
dealer
article
Topic 13
people group
would make
drug line
government
may
firearm use
Topic 14
team good
game
article write
year
line
subject would
player
Topic 15
be write
would article
get
line
time car
subject use
Topic 16
want line right
question
write
know
subject would
list article
Topic 17
chz
mh
tg eqtn
um_um
rlk
wwiz
ax
tq
ei_ei
Topic 18
wire scsi
drive
go bit get
ide line
chip second
Topic 19
write encryption
line
would could
chip
system
subject
use key