**CATARSI – PREPARAZIONE STRUMENTO WEB**

Questo documento vuole descrivere i passi da noi fatti tutte le volte in cui, in precedenza, abbiamo usato il Topic Modeling. Idealmente ogni passo potrebbe corrispondere a una casella da attivare nel processo on-line.

**1. Raccolta dei dati**

Mallet, lo strumento usato fino ad ora, vuole in input dati come file txt. Di conseguenza, abbiamo lavorato per ridurre a txt dati che fossero in altro formato. Queste le casistiche da noi incontrate:

1.1 Dati già in txt.

Mi è capitato di usare il TM per analizzare interviste. In quel caso ho semplicemente salvato il file della sbobinatura in txt.

1.2. HTML

Più interessante il caso dell’html, perché molto spesso ci siamo trovati ad analizzare testi che fossero su pagine internet (dico genericamente html, ma intendo pagine web). In questo caso, nel caso avessimo “pochi” dati, abbiamo proceduto facendo un bel copia/incolla manuale: ogni fonte un diverso txt. Per un altro progetto. Un collega a Vienna ha pagato una persona che ha adattato uno scraper al nostro caso (scaricare articoli di giornale da un quotidiano). Lo scraper salvava direttamente i dati in txt, dando al file un nome parlante, del tipo FONTE\_ANNO\_ID.txt. Se esiste e non implica perdere molto tempo, si potrebbe pensare di aggiungere uno strumento di scraping al nostro tool?

1.3 PDF.

Quando abbiamo lavorato con articoli accademici, abbiamo avuto a che fare con dei PDF, più o meno belli. Il nostro obiettivo finale era salvare solo il contenuto dell’articolo, eliminando nomi autori e titolo del paper, funding, acknowledgements, nomi autori e affiliation, references. Abbiamo tenuto le keywords. Dal punto di vista pratico:

1. Per i PDF testo, abbiamo usato “Calibre”, che è un software free per la gestione della libreria di ebook. Lo abbiamo usato per trasformare i pdf in txt.
2. Per i PDF immagine abbiamo usato un plug in di ADOBE (free per 30 giorni) che fa riconoscimento OCR. Poi abbiamo salvato il file in TXT
3. Poi abbiamo dovuto compiere una serie di operazioni manuali sui file TXT. Per farlo abbiamo usato Notepad++ , un software free che è una versione potenziata del blocco note (e che viene spesso usato anche solo per togliere formattazioni varie dai file word). Di seguito l’elenco delle operazioni effettivamente fatte, segue nota per generalizzare
   1. Sostituito con [vuoto] in tutti i documenti queste parole
      * ach.sagepub.com
      * Downloaded from
      * Copyright © 2006 SAGE Publications (per diversi anni)
      * sagepub.co.uk/journalsPermissions.nav
      * © The Author(s) 2011 (per diversi anni)
      * Reprints and permissions:
      * This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. [lasciato **Funding,** che va tolto a mano di volta in volta.]
   2. Operazioni preliminari:
      * Aprire un paper a caso (non un editorial) e togliere con Notepad++ le stringhe comuni a tutti i file di quell’ anno. Ad esempio tutti i file del 2000 hanno stringa “Accounting History NS Vol 5, No 2 – 2000”. Varia solo il numero del Vol e del No.
      * Aprire Notepad++
      * Search
      * Find in files
      * Find what: la stringa che si vuole togliere. Ad esempio “Accounting History NS Vol 5, No 2 – 2000”.
      * Replace with: vuoto
      * Filters: \*.\*
      * Directory: il folder “AH txt partially cleaned 2015.10.07”
      * Click on replace in files
      * Nel nostro esempio poi aggiornare numero Vol e No per togliere tutte le varianti.
   3. Operazioni file per file
      * Togliere tutto quanto sta sopra la prima parola del titolo
      * Togliere DOI:.....
      * Togliere nome autore/i e affiliation/s
      * Appena sotto introduction di solito appare nome e address contacting author: togliere
      * Individuare il cognome dell’ autore che compare ogni due pagine.
        1. Evidenziarlo
        2. Copiarlo: Ctrl C
        3. Aprire funzione replace: Ctrl H
        4. Incollarlo: ctrl V
        5. Click on “Replace all”
      * Fare stessa procedura per versione short titolo paper nel caso ci sia
      * Scroll down fino ad Acknowledgement e togliere tutto fino a parola Funding compresa
      * LASCIARE “Note” e testo sottostante
      * Individuare “References” e toglierle
      * OCIO che dopo references certe volte ci sono Appendix: queste sono da lasciare.
   4. Tempo stimato per anno: 15 min.

**1.4 NOTA:**

Ho riportato tutta questa procedura per spiegare cosa abbiamo fatto, ma ci sono almeno due cose che in futuro farei diversamente: 1) Difficile avere a che fare con PDF immagine. Nel nostro caso dovevamo lavorare con tutti gli articoli pubblicati su una specifica rivista, ma in generale si cambia campione se i dati sono terribili. 2) Alcune di queste pulizie le gestirei tramite stopword list: tutte le parole ripetute si possono mettere in stopword list, così come i numeri. Alcuni campi vanno però eliminati e non saprei come farlo, se non manualmente: nomi degli autori, funding, references. Tutto questo pezzo, comunque, serve più a spiegare come abbiamo fatto, che a replicarlo fedelmente. La cosa che sarebbe utile avere nel tool è uno strumento che trasformi i PDF in TXT, quantomeno.

**2. Preparazione dei dati**

La preparazione dei dati, prima di girarli a mallet consiste sostanzialmente in due operazioni:

1. Rinominare il txt in maniera tale che dal titolo si capisca cosa è: FONTE\_ANNO\_ID.txt. Ovviamente in altri casi le informazioni possono essere diverse. Se avessimo una fonte sola, ma una granularità maggiore, potrebbe essere ANNO\_MESE\_ID. Queste informazioni servono poi per l’interpretazione quando diventa necessario recuperare velocemente il file di partenza o capire distribuzione topic per fonte e/o anno.
2. Preparare una stopword list per “pulire” i file attraverso mallet. Ogni contesto, e quindi ogni progetto, ha una sua stopword list specifica. Anche la stopword list è in txt.
3. Una cosa che non serve per mallet ma diventa molto utile per interpretazione è un file excel che abbiamo preparato i metadati di ogni documento. Per ogni article ID erano riportate informazioni tipo: titolo, numero autori, gender, affiliation, country. Sarebbe bello se lo strumento potesse importare dati di questo tipo per permettere query del tipo: mostrami se di topic 8 parlano di più autori anglosassoni o indiani.

**3. Analisi dei dati con mallet**

Abbiamo analizzato i dati con Mallet. Di seguito i due comandi per 1) importare i dati su Mallet e per 2) analizzarli. Questo non per spiegare il linguaggio di Mallet, ma per spiegare input e output delle due fasi.

3.1 Importare i dati su mallet

bin\mallet import-dir --input sample-data\FBat --output FBat --keep-sequence --stoplist-file c:\mallet\stoplist\_file\_at.txt

bin\mallet import-dir 🡪 comando che indica l’operazione di importazione

--input sample-data\FBat 🡪 indica in quale cartella si trovano i fie txt

--output FBat 🡪 definisce il file mallet in cui tutti i txt sono importati

--keep-sequence 🡪 analizza I file in ordine (non troppo importante, sono possibili pure altre opzioni)

--stoplist-file c:\mallet\stoplist\_file\_at.txt 🡪 localizzazione della stopword list

3.2 Analizzare i dati

bin\mallet train-topics --input FBita --num-topics 20 --num-top-words 100 --num-iterations 3000 --output-state FBita20.gz --output-topic-keys FBita20-keys.txt --output-doc-topics FBita20-composition.txt --word-topic-counts-file counts-file-FBita20.txt --output-model FBita20model --diagnostics-file Fbita20.xml

Spiegazione:

bin\mallet train-topics 🡪 invoco train topics

--input FBita 🡪 file input

--num-topics 20 🡪 numero di topic che si vuole creare in output

--num-top-words 100 🡪 numero di parole che si vuole visualizzare per ogni topic

--num-iterations 3000 🡪 numero iterazioni

--output-state FBita20.gz 🡪 Gibbs sampling: dettaglio sulla codifica di ogni parola ne corpus.

--output-topic-keys FBita20-keys.txt 🡪 Definisce il nome del file txt su cui saranno salvate le top words per ogni topic e i parametri di dirilecht

--output-doc-topics FBita20-composition.txt 🡪 Mi fornisce il dettaglio di come ogni topic costituisca in diversa percentuale ogno documento orginario

--word-topic-counts-file counts-file-FBita20.txt 🡪 rappresentazione sparsa degli assignement word-topic

--output-model FBita20model 🡪 file in cui è scritto il binary topic model alla fine dell’esecuzione. Questo dovrebbe essere riutilizzabile per provare altri documenti sullo stesso topic model, ma non siamo capaci

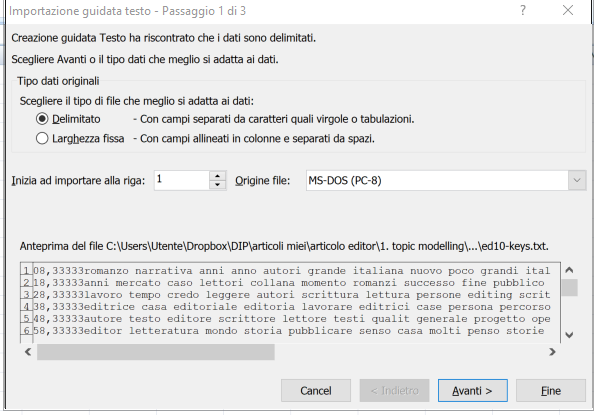
--diagnostics-file Fbita20.xml 🡪 file in cui ci dà misure diagnostiche sulla qualità del modello

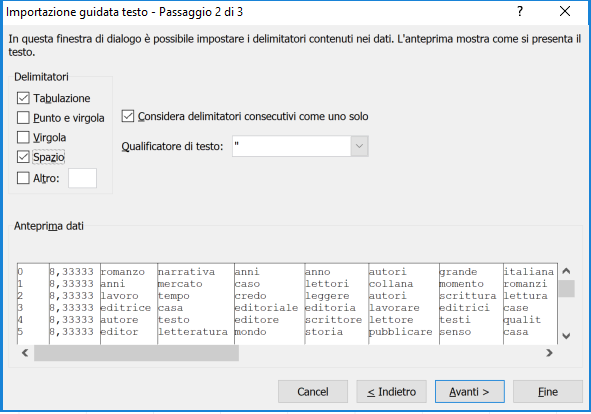
**4 Interpretazione dei risultati**

Il passaggio più importante è quello dell’interpretazione dei risultati. Qui diventa molto importante poter visualizzare i risultati in modi diversi, possibilmente non dovendo rifare a mano i passaggi tutte le volte. Noi per questa fase usiamo soprattutto i file **keys** e **composition**.

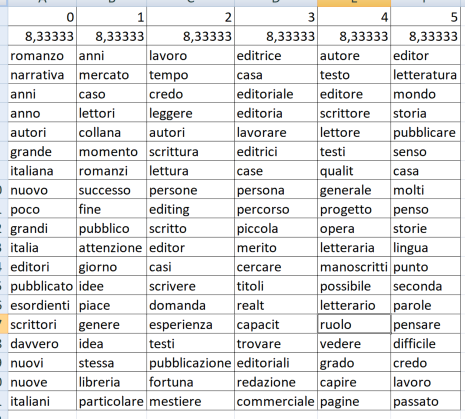
4.1 File Keys

Si tratta di fare dati -> da testo e importare il file keys.txt. Esempio





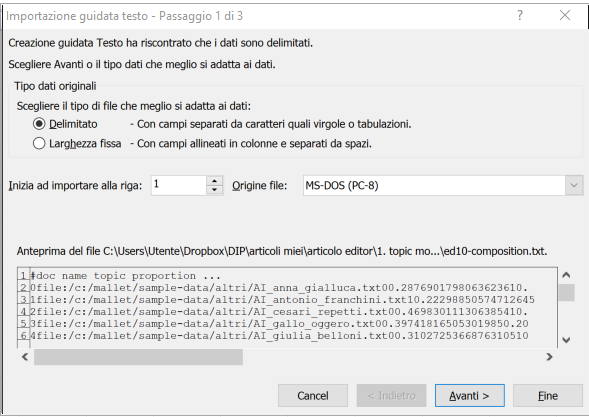
Poi si traspongono i risultati e si ottiene una tabella in cui ogni colonna è un topic. La prima cella è in numero progressivo del topic, la seconda è un parametro di Dirilecht, poi seguono le parole (quante ne avevamo richieste) in ordine dalla più importante alla meno importante

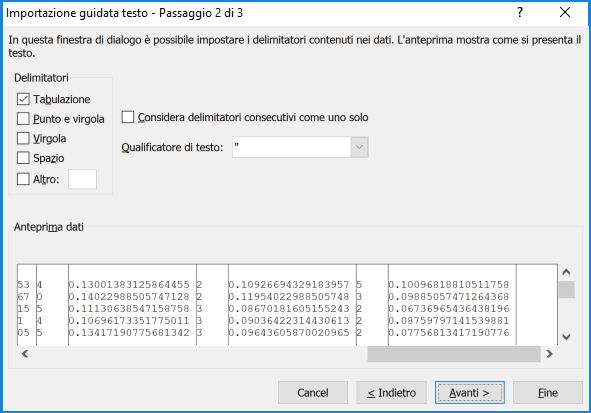


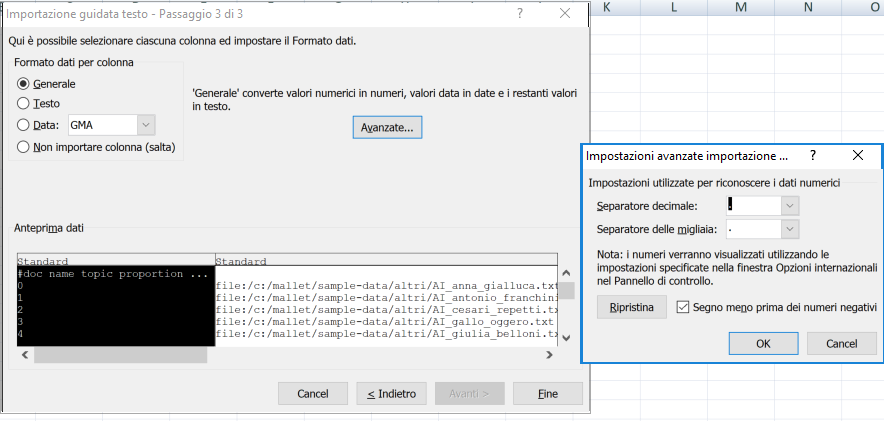
Questa cosa si usa per interpretare il senso di ogni topic.

4.2 File Composition

Di nuovo si importa da dati -> da testo e si importa il file con la composizione

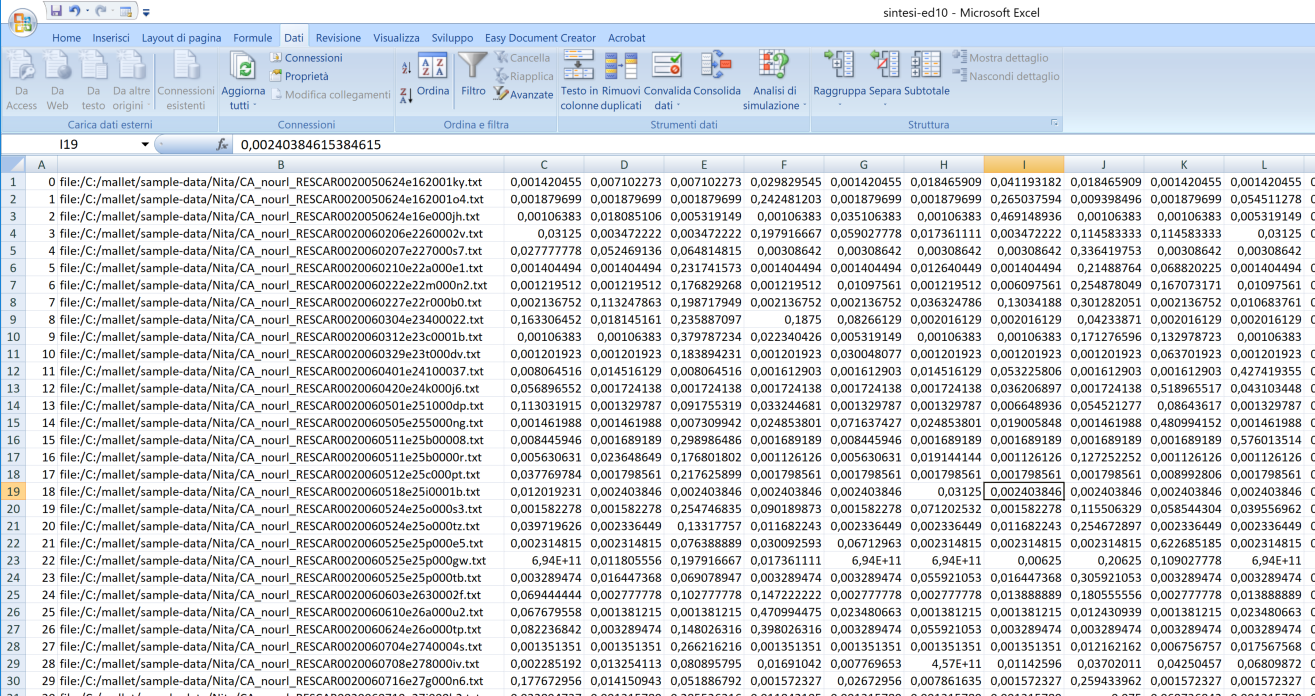






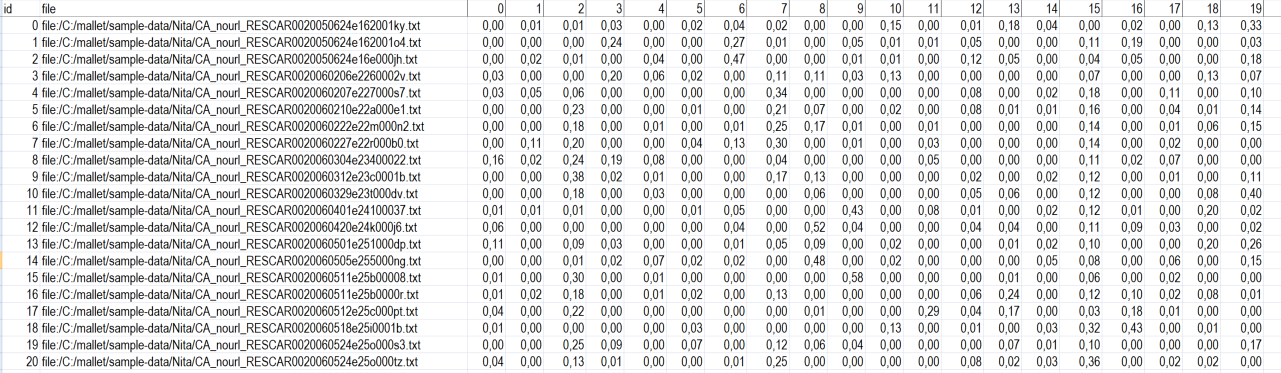
Siccome excel è in italiano e Mallet è americano devo anche spiegare qual è il separatore decimale, se no prende i numeri per stringhe di testo.

Il risultato ha questa forma



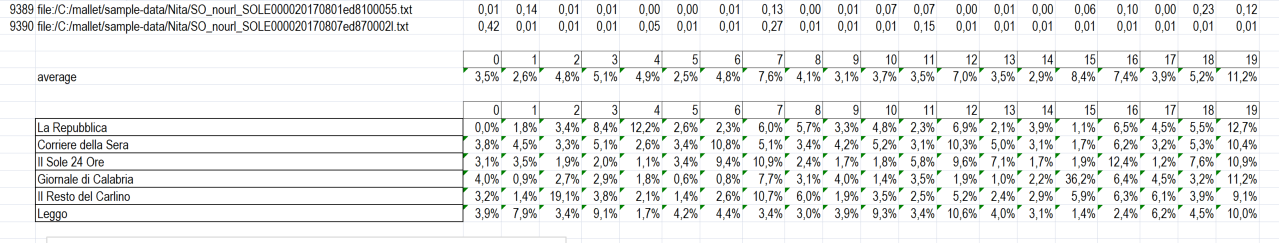
Ogni riga corrisponde a un articolo. La prima colonna è un ID, la seconda colonna individua il eprcorso originario del file (con il nome), dalla terza colonna in poi ho la composizione in topic di ogni testo. Per cui, ad esempio, il primo file è costituito dal topic 0 per lo 0,14%, dal topic 1 per lo 0,7%, dal topic 3 per lo 0,7%, dal topic4 per il 2,98% e così via.

Purtroppo importando questo file gli esponenziali negativi diventano degli esponenziali positivi, per qualche motivo che non so. Ad esempio la casella C23 dovrebbe essere 6,94\*e-11, ma diventa 6,94\*e11. Questo genere di problemi lo risolvo con operazioni di excel. Poi riformatto la tabella e alla fine ho una cosa del genere:



(ho messo le prime 20 righe su 9390).

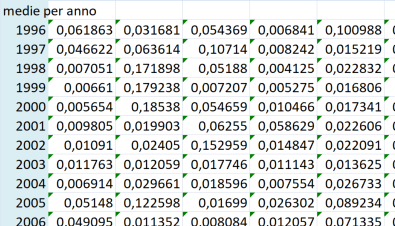
Da qui ci inventiamo modi per visualizzare i risultati. Nel caso di cui sopra, ad esempio, lavoravo con articoli di giornale, presi da 6 fonti diverse. Per cui ho calcolato l’uso medio di ciascun topic sul campione totale e fonte per fonte.



Tabulandolo

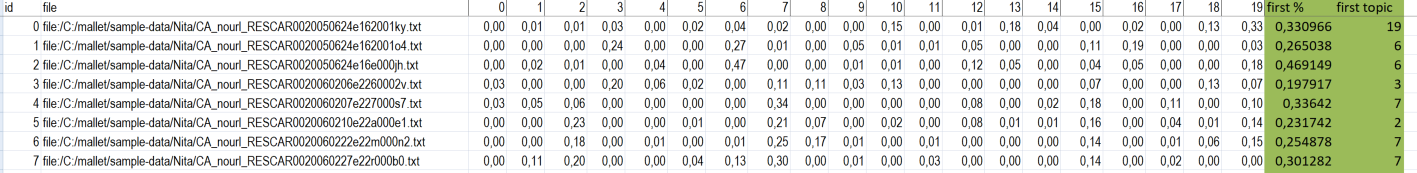
Qui la cosa importante è che serve fare statistiche descrittive su caratteristiche dei file. In questo caso si distinguevano solo per fonte, ma in altri casi si può calcolare l’uso dei topic per anno, per costruire trend nel tempo. La cosa importante per noi è che si tenga traccia di queste variabili nel nome del file, per poter usare il nome per ordinare le caselle di excel, in modo da calcolare medie diverse.

Ad esempio, in un altro progetto avevamo le medie per anno per ogni topic, che poi abbiamo graficato

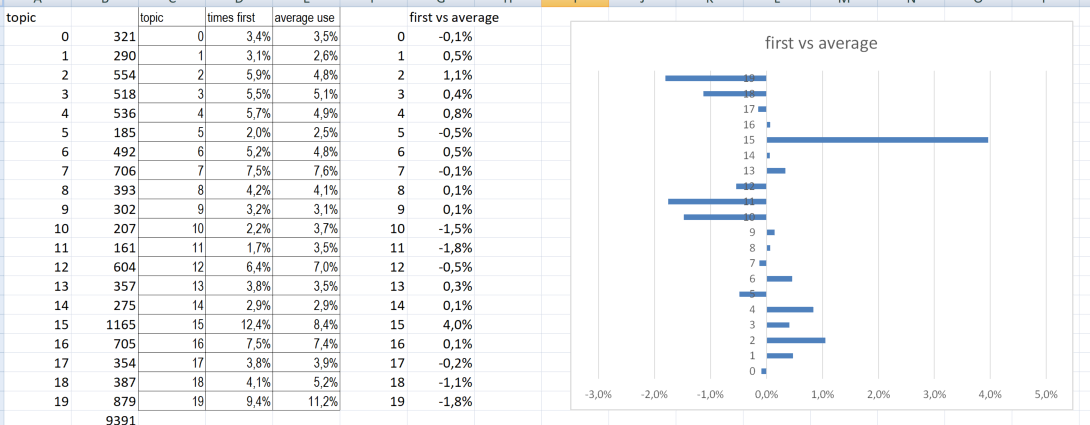


Il punto cruciale è che le righe su cui calcolare la media variano da progetto a progetto. L’ideale per visualizzare sarebbe avere uno strumento che calcoli medie dando la possibilità di selezionare le righe su cui calcolarle.

Un’altra cosa che abbiamo calcolato, a partire dal dato excel, è, per ogni riga, qual è il topic più importante e qual è il suo uso in quel topic.



Questa cosa permette di capire come viene usato un topic. Ad esmepio un topic che costituisce in media il 4% degli articoli, ma che nell’8% dei casi è il topic più importante per un articolo è un topic che viene usato solo in articoli che parlano proprio di quel topic. Altre volte abbiamo calcolato, sempre tramite excel, i tre topic principali di ogni articolo, per poi ragionare su quante volte rispetto al totale un topic era uno dei primi tre. Oppure si può tabulare “times firts vs average use”



Qui non saprei bene cosa chiedere a un tool automatico. Più che altro ti descrivo cosa abbiamo fatto noi, per capire che genere di analisi fare.

Diciamo che in un mondo perfetto sarebbe bello avere un software cui dici:

1. Calcolami il topic model con la soluzione a 20 topic e a 21
2. Calcola le media d’uso dei topic per entrambe le soluzioni su tutto il campione
3. Calcola la media d’uso dei topic per entrambe le soluzioni su alcuni documenti selezionati

Potremmo poi fare i grafici a mano, ma questo sarebbe già un bel passo avanti

**5. Query frequenti**

Per completare quanto scritto sopra, riportiamo sotto le query a cui vorremmo che il tool rispondesse. Sono tutti passaggi per l’interpretazione che al momento vengono fatti muovendosi tra diversi fogli di file excel

* Accorpa il valore medio di topic X, Y, Z e quello di topic A, B, C e confrontali
* Mostrami topic 9 per fonte/anno/affiliation autori
* Mostrami trend da anno 1 a anno 10 per topic 4 e 7 o per la somma di topic 5+6 e 7+8
* Mostrami i 10 documenti in cui topic 7 è maggiormente presente. Qui sarebbe l’ideale un link al documento o magari poter visualizzare i metadati del documento
* Mostrami i documenti dove topic 8 < 3%
* Mostrami i documenti dove sia topic 1 che topic 10 sono > di 15%

**6. Temi aperti**

Finisco con due questioni aperte.

5.1 Altre analisi con TM

Ci sono altre analisi, che si possono fare col Topic Modeling, e che noi non sappiamo fare (ad esempio paragrafo 6.1 paper Di Maggio). Può essere che pacchetti già esistenti che applichino LDA possano fare alcune di queste cose. Teniamolo buono per il futuro.

5.2 Scelta del numero ottimo di Topics

Ci sono due scuole in letteratura: chi dice che non si può trovare un numero ottimo di topic per definizione e chi cerca di calcolare il numero ottimo di topic con modelli matematici. Ora, nonostante chi ha scritto gli articoli più importanti è della prima idea, molti reviewer sono della seconda. Sarebbe quindi molto utile pensare anche a un tool che calcola la bontà della soluzione. Però ci sto lavorando ora e non so dirvi come fare, devo prima leggere cose. Teniamolo come cosa aperta per il futuro!