Ingegneria Informatica e dell'Automazione

Università degli Studi di Perugia Facoltà di Ingegneria

Candidato: Enrico Spataro

Relatore: Prof.ssa Carla Binucci Relatore: Prof. Walter Didimo

19.02.2016



Contenuti

- 1 Introduzione
- 2 Creazione word cloud Word cloud statica Word cloud dinamica
- 3 Risultati sperimentali
- 4 Conclusioni



Obiettivi

Mostrare l'evoluzione di un testo tramite:

- creazione word cloud semantiche ad intervalli regolari;
- animazioni tramite morphing.



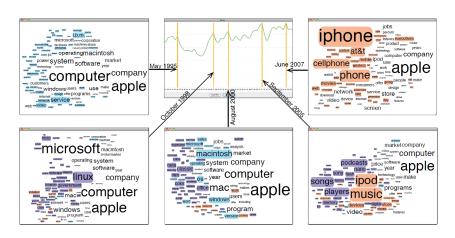
Cos'è una word cloud?



Word cloud semantiche (Kobourov et al., 2014)

interrogator ESP kind mistakes identification GUESS Christmas Lovelace Babbage clairvoyance telepathy criterion satisfy pupil fault discrete-state Turing solipsist analogy infinite satisfactorily speed suitable induction child programme random machine wheel surprise fallacies suppose Yes conjectures wheel machine wheel mere wheel mere wheel surprise fallacies suppose Yes conjectures wheel mere wheel mere wheel surprise fallacies suppose Yes conjectures and the surprise fallacies suppose Yes conjectures wheel mere wheel surprise fallacies suppose Yes conjectures and the surprise fallacies suppo game behaviour man argument quite clicks homework answer wrong try instance convenient packets store instruction obey

computer



Morphing (Chi et al., 2015)

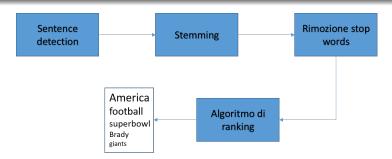


Word cloud statica

- Estrazione parole chiave
- Calcolo similarità
- Clustering
- Disegno word cloud



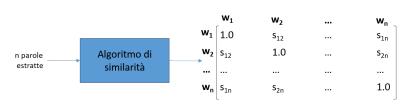
- Estrazione parole chiave
- Calcolo similarità
- Clustering
- Disegno word cloud





Word cloud statica

- Estrazione parole chiave
- Calcolo similarità
- Disegno word cloud



$$s_{ij} = s_{ji} \text{ , } s_{ij} \in [0, 1]$$

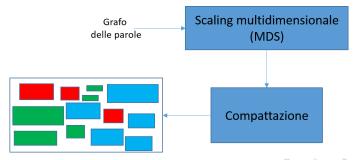


- 1 Estrazione parole chiave
- Calcolo similarità
- 3 Clustering
- 4 Disegno word cloud



Word cloud statica

- Estrazione parole chiave
- Calcolo similarità
- Clustering
- 4 Disegno word cloud





•000000

Obiettivo: create K word cloud dal testo, si vuole passare da una word cloud alla successiva in modo graduale tramite tecniche di morphing.



000000

Morphing

Caratteristiche

- Numero di frame
- Gestione stato delle parole



Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

- scomparire
- apparire
- rimanere nel layout (variando posizione)



Le parole, tra una word cloud e la successiva, possono:

0000000

- scomparire







Word cloud dinamica

Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

0000000

- scomparire
- apparire
- rimanere nel layout (variando posizione)





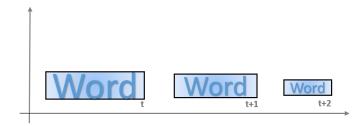


Stato delle parole

Le parole, tra una word cloud e la successiva, possono:

0000000

- scomparire
- apparire
- rimanere nel layout (variando posizione)

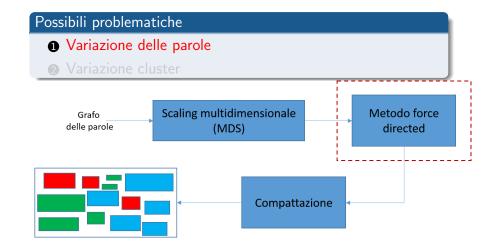


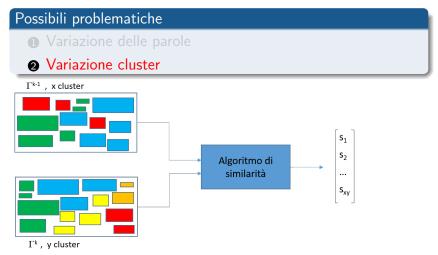


Possibili problematiche

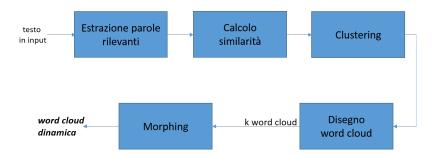
- Variazione posizione delle parole
- 2 Variazione dei cluster







Architettura



Creazione word cloud

Word cloud dinamica

Risultati sperimentali

Test suite:

- 200 discorsi estratti (file .txt) dal set di conferenze annuali TED (Technology Entertainment Design).
- Lunghezza media 17/18 minuti.
- 4 campionamenti per ogni testo.
- Parole estratte: 20.40.60



Metriche adottate:

- Combination metric $\Rightarrow \nu = \frac{1}{K} \sum_{k=1}^{K} (\alpha S^k + \beta \vartheta^{k,k-1})$
 - Distortion metric $\Rightarrow S^k = \frac{\sum_{ij} c_{ij} s_{ij}}{\sum_{ij} s_{ij}}$
 - Coherence metric $\Rightarrow \vartheta^{k,k-1} = 1 \frac{\sum\limits_{i=1}^p \sigma(w_i^k,w_i^{k-1})}{pD}$
- Space metric $\Rightarrow \gamma = 1 \frac{\mu}{\varphi}$
- Tempo d'esecuzione



Risultati sperimentali

Risultati sperimentali

Algoritmi utilizzati:

Ranking

Term Frequency

TF-IDF

LexRank

Similarity

Jaccard

Cosine

Extended Jaccard

Layout

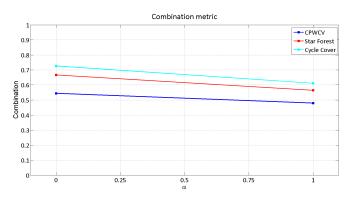
CPWCV

Star Forest

Cycle Cover



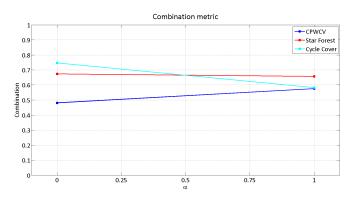
Combination metric



Parole estratte: 20

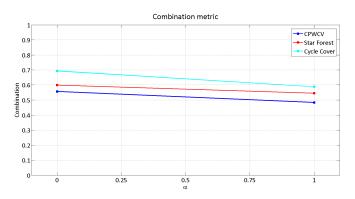


Combination metric



Parole estratte: 40

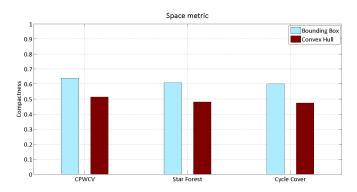




Parole estratte: 60



Space metric



Parole estratte: 40



Tempo d'esecuzione

- Configurazione caso peggiore: TFIDF, Jaccard Similarity e Star Forest
- Configurazione caso migliore: Term Frequency, Cosine Similarity e CPWCV

Parole estratte	Caso peggiore	Caso migliore
20	2.65 <i>s</i>	0.77 <i>s</i>
40	3.01 <i>s</i>	0.94 <i>s</i>
60	3.66 <i>s</i>	1.25 <i>s</i>

Conclusioni

- Nuovo approccio per la generazione di word cloud dinamiche.
- Buoni risultati dagli algoritmi di layout utilizzati.
- Tempo d'esecuzione soddisfacente.

Sviluppi futuri:

- Utilizzo di tecniche più sofisticate di Text Mining (e.g. POS Tagging).
- Definizione di metriche che tengano conto dell'interfaccia grafica + user study.



Grazie per l'attenzione