CORSO: Text mining and document analysis

DOCENTI: Alessio Farcomeni (Ph.D., 2004); Marco Stefanucci (Ph.D., 2018)
EMAIL: alessio.farcomeni@uniroma2.it
marco.stefanucci@uniroma2.it
PAGINE WEB: https://economia.uniroma2.it/faculty/563/farcomeni-alessio
https://economia.uniroma2.it/faculty/725/stefanucci-marco

DESCRIZIONE DEL CORSO

Il corso discute di alcune tecniche di analisi di dati testuali. Molto spazio verrà dato alla costruzione del data set a partire da un corpus di documenti. Verranno utilizzati casi di studio basati su testi brevi (ad esempio, da social media) e meno brevi (ad esempio, articoli di giornali). Verranno introdotte tecniche di descrizione del corpus, di segmentazione dei testi e di sentiment analysis.

OBIETTIVI DI APPRENDIMENTO

- ✓ Abilità di ottenere un data set a partire da un corpus di testi, facendo le scelte in fase di preprocessing più adatte agli obiettivi del progetto
- ✓ Abilità di sintetizzare le informazioni in un corpus testuale.
- ✓ Abilità di identificare gli argomenti in un corpus testuale
- ✓ Abilità di sintetizzare il sentiment in un corpus, sia in maniera supervisionata che non supervisionata

METODOLOGIA

L'enfasi è sui principi e sulle tecniche statistiche specifiche. Ciascun metodo è introdotto tramite esempi e approfondito da un punto di vista tecnico. Una base di statistica matematica è necessaria, ma le derivazioni verranno ridotte al minimo indispensabile. Le metodologie sono discusse da un punto di vista teorico e pratico, con forte enfasi sulla parte pratica.
Vengono descritte le definizioni, assunzioni, proprietà, implementazione, e interpretazione di ciascuna metodologia. L'intero corso è basato sul software *R.*

VALUTAZIONE

Esame scritto, basato su domande chiuse (con la possibilità di sporadiche domande aperte). L'esame verterà sugli aspetti di specificazione e interpretativi delle metodologie discusse. Alcune domande riporteranno anche codice o output ottenuto dal software *R*, su cui verterà lo specifico quiz.

PROGRAMMA

- 1. Introduzione alla analisi testuale ed overview del corso
- 2. Gestione del testo: tokenization, N-grams. Cenni a named entity recognition e part of speech tagging
- 3. Analisi testuale e strumenti descrittivi
- 4 Topic modeling. Latent Dirichlet allocation
  4.1 Cenni a rappresentazioni dipendenti dal contesto, word embeddings, BERT.
- 5. Sentiment analysis supervisionata e non-supervisionata

Se il tempo lo permette, possono essere discussi altri argomenti aggiuntivi o propedeutici alla comprensione dei contenuti del corso.

LIBRI DI TESTO

Silve J. and Robinson D. (2017) *Text Mining with R*, O'Reilly Media, Inc.
(https://www.tidytextmining.com/)