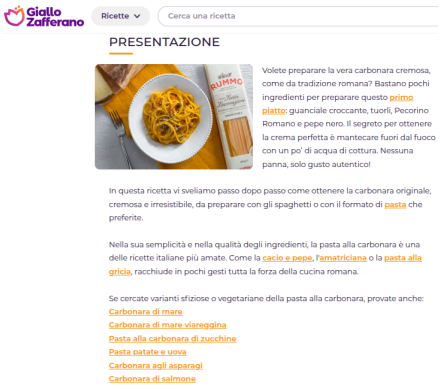
Homework 2

Link a progetto Github: https://github.com/giorgiomelch/Ingegneria-dei-Dati_h2.

I file da indicizzare sono stati raccolti da Giallo Zafferano, salvati in formato .txt. Il filename è il nome della ricetta, mentre il contenuto del file è la presentazione del piatto.

Esempio:



- Filename: “Spaghetti alla cabonara”
- Contenuto: “Volete preparare la vera carbonara cremosa, come da tradizione romana? Bastano pochi ingredienti per preparare questo primo piatto: guanciale croccante, tuorli, Pecorino Romano e pepe nero. Il segreto per ottenere la crema perfetta è mantecare fuori dal fuoco con un po’ di acqua di cottura. Nessuna panna, solo gusto autentico! In questa ricetta vi sveliamo passo dopo passo come ottenere la carbonara originale, cremosa e irresistibile, da preparare con gli spaghetti o con il formato di pasta che preferite. Nella sua semplicità e nella qualità degli ingredienti, la pasta alla carbonara è una delle ricette italiane più amate. Come la cacio e pepe, l'amatriciana o la pasta alla gricia, racchiude in pochi gesti tutta la forza della cucina romana.”

Analyzer scelti e motivazioni:

- **PerFieldAnalyzerWrapper**: è stato utilizzato per assegnare un analizzatore diverso a ciascun campo del documento indicizzato. In questo modo è possibile trattare in maniera distinta il contenuto testuale e il nome del file.
- **SimpleAnalyzer** per il campo "**filename**": il nomi dei file sono stringhe brevi e semplici. Consente una tokenizzazione semplice, che suddivide il testo in token

alfanumerici e converte tutto in minuscolo, senza applicare rimozione di stopword o tecniche linguistiche.
- **ItalianAnalyzer** per il campo "**content**": gestisce stemming, stopword italiane e tokenizzazione, adatto per il dominio di ricette italiane permettendo di collegare varianti morfologiche di uno stesso termine (es "cottura", "cotture", "cuocere").

Numero di file indicizzati e tempi:

Sono stati indicizzati **51 file** con parole medie per il titolo pari a 3,68 , per contenuto 145.35.
Tempo di indicizzazione: 723 ms.

Query usate per testare il sistema

nome pasta
nome alla
contenuto zafferano
contenuto carne -secondo
contenuto pesce +primo
contenuto buono +carne
contenuto zafferano NOT pasta
contenuto -pasta +carne
contenuto buon*
contenuto +pasta +guanciale -tuorlo -uovo