



Race for the White House

With(Out) feelings

Sviluppata da:

Riccardo Giomi **495996**

Edoardo Carmazzi **491246**

Progetto a scopo didattico per l'esame di Basi di Dati
e Laboratorio Web

- Il progetto..... pag. 3
 - Reddit.com..... pag. 3

- Utilizzare Reddit..... pag. 4

- Sentiment Analysis..... pag. 5
 - Stato dell'arte..... pag. 5

- Costruzione e architettura applicazione..... pag. 7

- Struttura Grafici.....pag. 9

- Problematiche e limiti..... pag. 11
Possibili sviluppi e conclusioni

Il progetto

Uno dei temi più caldi e discussi del 2016, è sicuramente quello delle elezioni presidenziali negli Stati Uniti d'America. Il successore di Barack Obama verrà eletto l'8 novembre 2016 e dovrà guadagnarsi il consenso degli elettori nei 50 stati del territorio Americano.

Giugno 2016

Dopo le primarie, la lista dei candidati alla presidenza degli Stati Uniti si riduce a due: Hillary Clinton per i Democratici, e Donald Trump per il Partito Repubblicano.

L'obiettivo della nostra applicazione è quello di stimare la popolarità dei candidati sul web, effettuando un'analisi dei commenti ai post presenti su Reddit, il sito di social news più visitato in America.

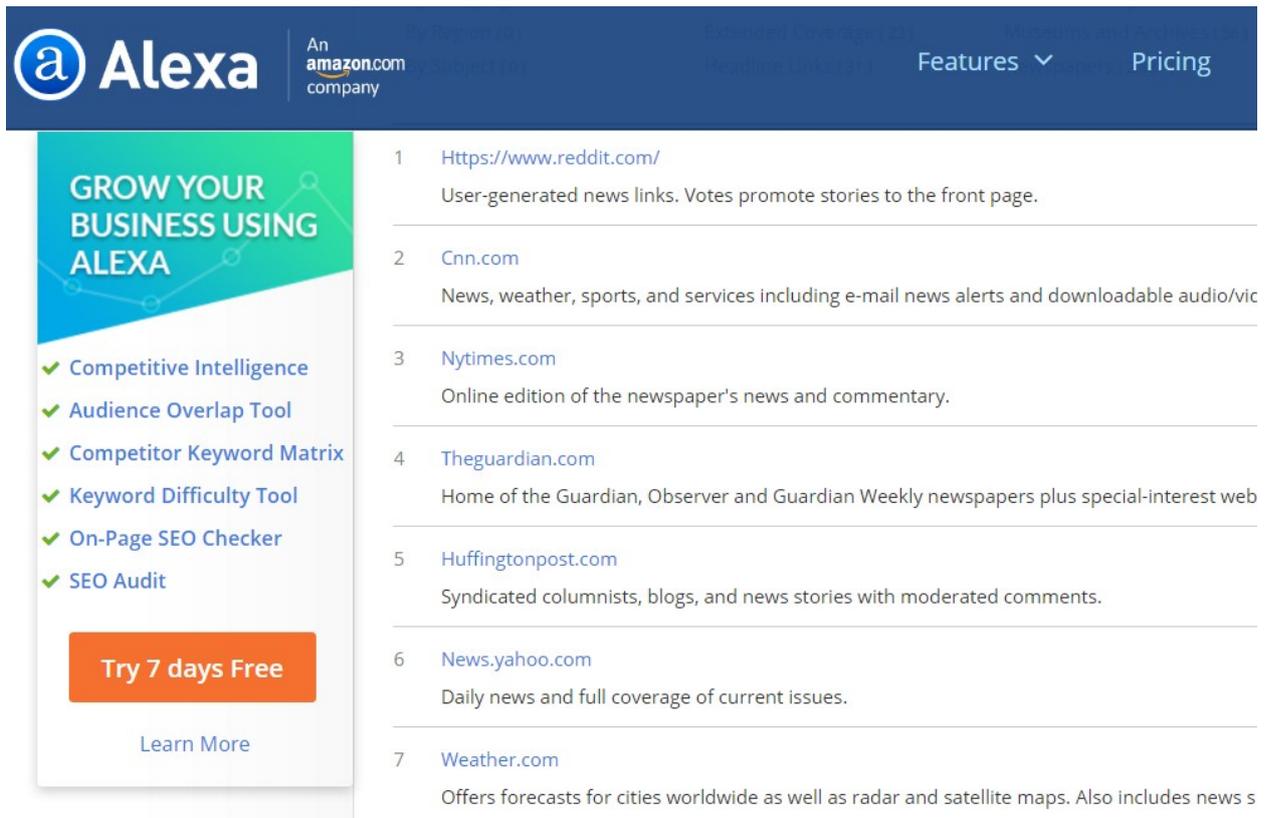
Reddit: the front page of internet

Reddit, da "read + edit" ("leggi e "modifica") è diventato negli ultimi anni molto popolare negli States, classificandosi al 9° posto tra i siti più visitati e al momento si colloca tra Twitter e Netflix. Vanta inoltre il primato delle visite nella categoria news.

The screenshot shows the Alexa website interface. At the top, there is a navigation bar with the Alexa logo, the text 'An amazon.com company', and links for 'Features' and 'Pricing'. Below the navigation bar, there is a sidebar on the left with a green and blue background. The sidebar contains the text 'GROW YOUR BUSINESS USING ALEXA' and a list of tools: 'Competitive Intelligence', 'Audience Overlap Tool', 'Competitor Keyword Matrix', 'Keyword Difficulty Tool', 'On-Page SEO Checker', and 'SEO Audit'. Below the list of tools is a button that says 'Try 7 days Free' and a link that says 'Learn More'. To the right of the sidebar is a list of top websites in the USA. The list is numbered 4 through 10 and includes the following entries: 4. Amazon.com (Amazon.com seeks to be Earth's most customer-centric company, where customers can find a...), 5. Yahoo.com (A major internet portal and service provider offering search results, customizable content, cha...), 6. Wikipedia.org (A free encyclopedia built collaboratively using wiki software. (Creative Commons Attribution-Sh...), 7. Ebay.com (International person to person auction site, with products sorted into categories.), 8. Twitter.com (Social networking and microblogging service utilising instant messaging, SMS or a web interfac...), 9. Reddit.com (User-generated news links. Votes promote stories to the front page.), and 10. Netflix.com (Flat monthly fee streaming TV and movies service).

Categoria TopSites/USA. Fonte: (<http://www.alexa.com/topsites/countries/US>)

Categoria News. Fonte (<http://www.alexa.com/topsites/category/Top/News>)



The screenshot shows the Alexa website interface. At the top, there is a dark blue header with the Alexa logo, the text "An amazon.com company", and navigation links for "Features" and "Pricing". Below the header, on the left, is a white sidebar with a green and blue background. It features the text "GROW YOUR BUSINESS USING ALEXA" and a list of tools with green checkmarks: "Competitive Intelligence", "Audience Overlap Tool", "Competitor Keyword Matrix", "Keyword Difficulty Tool", "On-Page SEO Checker", and "SEO Audit". Below the list is an orange button that says "Try 7 days Free" and a "Learn More" link. To the right of the sidebar is a list of top news sites, numbered 1 through 7, with their respective descriptions.

Rank	Website	Description
1	https://www.reddit.com/	User-generated news links. Votes promote stories to the front page.
2	Cnn.com	News, weather, sports, and services including e-mail news alerts and downloadable audio/vic
3	Nytimes.com	Online edition of the newspaper's news and commentary.
4	Theguardian.com	Home of the Guardian, Observer and Guardian Weekly newspapers plus special-interest web
5	Huffingtonpost.com	Syndicated columnists, blogs, and news stories with moderated comments.
6	News.yahoo.com	Daily news and full coverage of current issues.
7	Weather.com	Offers forecasts for cities worldwide as well as radar and satellite maps. Also includes news s

Come funziona

Le pagine di Reddit sono divise in sottocategorie o **subreddit**, spazi di discussione moderati da admin, che dispongono regole e consigli da seguire all'interno della community. Le modalità di navigazione sono più o meno quelle di un grosso forum.

Tra i subreddit più popolari ci sono i cosiddetti "**IAmA**", inglese per "Io sono un", ovvero sezioni dove l'admin si presenta e invita gli iscritti a rivolgergli qualsiasi tipo di domanda a cui dovrà rispondere secondo la logica "**AMA**" ("Ask me anything"). Questa modalità ha raggiunto il picco di popolarità nell'agosto 2012, quando il presidente Barack Obama ha aperto una discussione **AMA** mettendo in crisi i server per l'enorme traffico dei partecipanti.

Su ogni subreddit è possibile postare link inerenti al tema della pagina, ogni post riceve un voto (positivo o negativo) fino a raggiungere un determinato punteggio, dato dalla sottrazione dei voti negativi a quelli positivi. Lo stesso vale per i commenti su ogni post. Durante la navigazione di un subreddit, l'utente può decidere se ordinare i post da visualizzare per *più recenti*, *più votati*, *più discussi* e *in crescita*.

Sentiment Analysis

L'Analisi del sentiment o "opinion mining" è la maniera a cui ci si riferisce all'uso dell'elaborazione del linguaggio naturale, analisi testuale e linguistica computazionale per identificare ed estrarre informazioni soggettive da diverse fonti. L'analisi del sentiment è ampiamente applicata per analizzare social media per una varietà di applicazioni, dal marketing al servizio clienti.

Questo tipo di operazione è in grado di fornire una gradazione di polarità di un contenuto testuale che va da "-1.0"(negativo) a "1.0"(positivo), lo "0" rappresenta la neutralità.

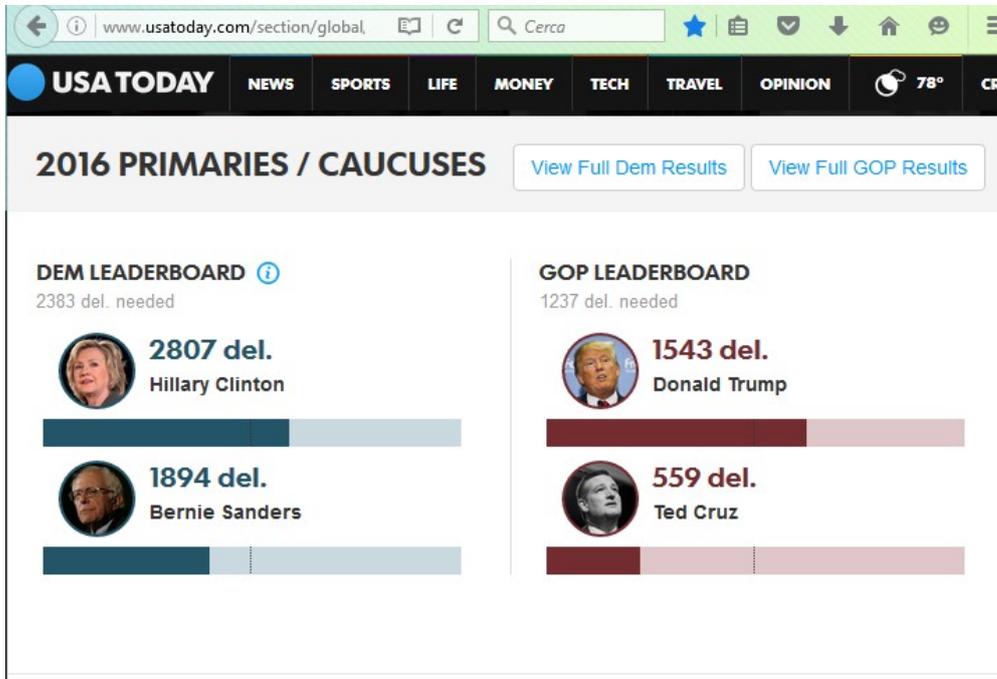
Stato dell'arte

Il nostro progetto tratta le elezioni USA 2016, un argomento decisamente attuale. Inoltre Reddit non è ancora così popolare in Italia e difficilmente potrebbe esistere un'applicazione analoga.

Da una ricerca sul web, non abbiamo riscontrato nessuna corrispondenza tra argomento trattato, tipologia di analisi effettuata e mezzo di comunicazione scelto.

Esistono però pagine che calcolano in tempo reale chi è in vantaggio secondo i sondaggi. Due esempi qua sotto:

- fonte: <http://www.usatoday.com/section/global/elections-2016/>



- fonte: <http://www.270towin.com/>



Costruzione e architettura applicazione

Il database è stato creato tramite la libreria “**PRAW**”, (Python Reddit API Wrapper), una libreria python per l'estrazione di dati.

Importandola, abbiamo creato un insieme di funzioni che ci hanno permesso di raccogliere dati relativi a post e commenti.

In particolare:

- il contenuto del commento,
- stato di provenienza dell'autore del commento,
- stato di provenienza dell'autore del post
- punteggio del commento,
- punteggio del post,
- ID del post,
- ID del commento

Una volta ottenuto e visualizzato in un formato .XML , abbiamo convertito l'output in .CSV e creato il database su *PHPmyadmin*.

```
import praw

user_agent = ("giomarawo10")

r = praw.Reddit(user_agent = user_agent)
subreddit = r.get_subreddit("The_Donald")
already_done = set()
print "<Structure>"

for post in subreddit.get_new(limit=1000):
    submission = r.get_submission(submission_id=post.id)

    forest_comments = submission.comments
    flat_comments = praw.helpers.flatten_tree(submission.comments)

    for comment in flat_comments:

        if not isinstance(comment, praw.objects.MoreComments) and comment.id not in already_done:

            print "<CommentBody>", comment.body, "</CommentBody>"
            print "<CommentAuthorState>", comment.author_flair_text, "</CommentAuthorState>"
            print "<CommentScore>", comment.score, "</CommentScore>"
            print "<PostScore>", post.score, "</PostScore>"
            print "<PostID>", post.id, "</PostID>"
            print "<PostAuthorState>", post.author_flair_css_class, "</PostAuthorState>"
            print "-----fine commento-----\n"

print "_____fine post_____ \n"
```

Per quanto riguarda il sentiment analysis, abbiamo utilizzato **TextBlob**, una libreria Python per l'elaborazione di dati testuali. Fornisce una serie di funzioni per calcolare polarità, soggettività, classificazione e traduzione di un testo.

```
get_comments_from_post.py x sentiment.py x
1 import csv
2 from textblob import TextBlob
3
4 import sys
5 reload(sys)
6 sys.setdefaultencoding("utf-8")
7
8 print "<Struttura>"
9 with open('hillary_new_comments.csv', 'r') as f:
10     reader = csv.reader(f)
11     for row in reader:
12         sentence = row[0]
13         blob = TextBlob(sentence)
14         print "<Polarita>", blob.polarity, "</Polarita>"
15         print "<Subjectivity>", blob.subjectivity, "</Subjectivity>"
16         print "---Prossima analisi---"
17 print "</Struttura>"
```

Struttura grafici

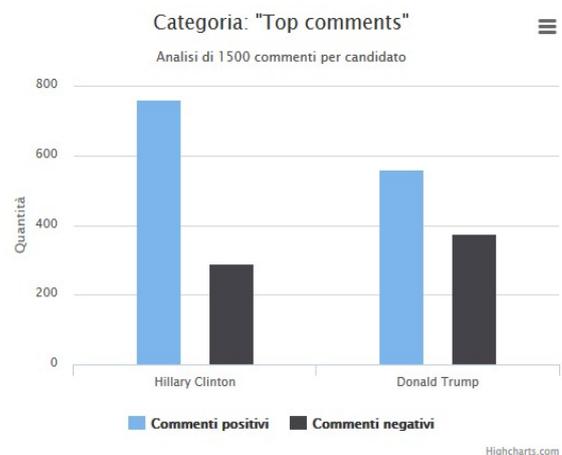
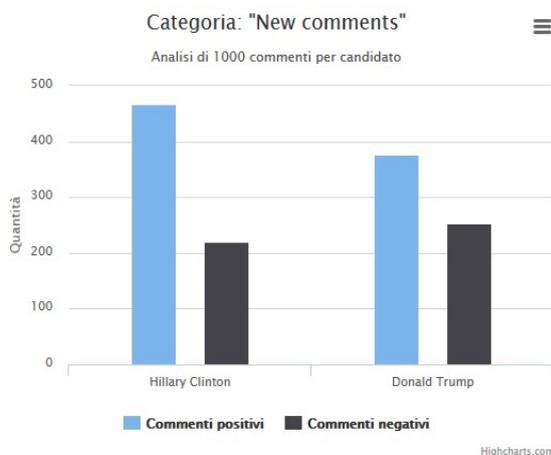
I linguaggi utilizzati per delineare la struttura e la grafica dell'applicazione sono HTML5 e CSS3. La parte dinamica è stata realizzata in Javascript con l'ausilio delle librerie jQuery e Bootstrap per le animazioni, Highcharts per i grafici a barre.

I grafici del sentiment analysis mostrano la quantità di commenti positivi/negativi per due subreddit: https://www.reddit.com/r/The_Donald/, e <https://www.reddit.com/r/hillaryclinton/>.

Entrambi sono i principali subreddit dedicati alla categoria in questione e non sono privati ma di libero accesso per tutti gli utenti di Reddit.

È interessante notare come nel grafico a destra relativo alla categoria "top comments" (commenti relativi ai post più votati), i commenti positivi e negativi per Donald Trump siano rispettivamente 550 e 380, su un totale di 1500 analizzato. La parte di dati mancante è dovuta a una grande quantità di commenti con polarità "0", classificati come *neutri* in fase di sentiment analysis, che abbiamo escluso a causa di una forte presenza di video e collegamenti ipertestuali.

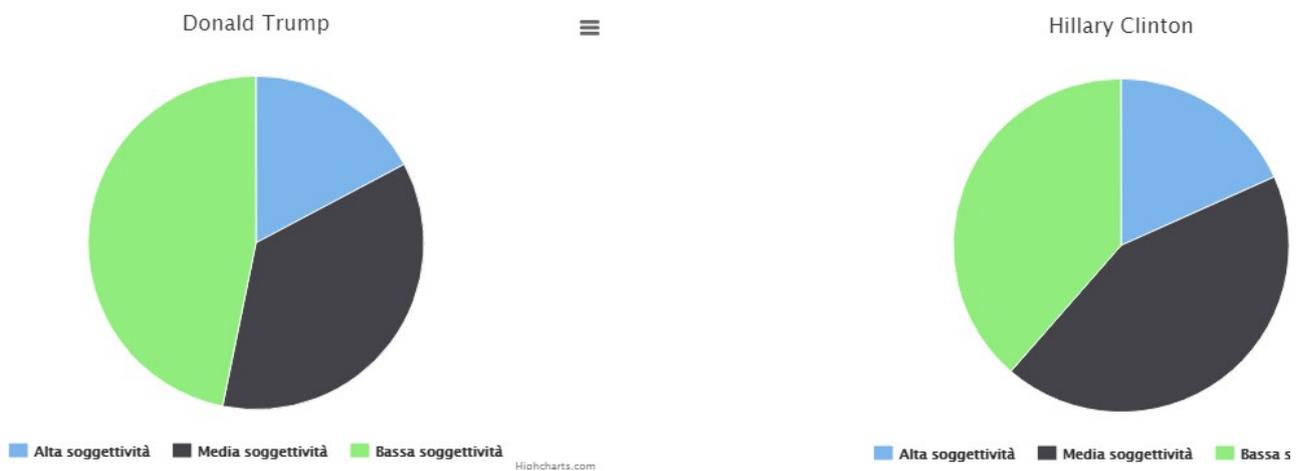
SENTIMENT ANALYSIS



I grafici della soggettività, derivano sempre dall'analisi del sentiment, e ci danno un'idea della percentuale di soggettività dei commenti.

Dai dati si evince che sulla parte azzurra (alta soggettività) si equivalgono i commenti di entrambi i candidati, mentre per Donald Trump spicca una bassa soggettività nei commenti (anche qua dovuta al fatto che la maggior parte sono link o contenuti multimediali).

SOGGETTIVITA DEI COMMENTI



Limiti e problematiche incontrate

L'estrazione e la pulizia dei dati ci sono risultate difficili da effettuare, poiché abbiamo dovuto capire come utilizzare gli attributi e le classi giuste per gli oggetti che volevamo.

In più Reddit presenta dei limitazioni, in quanto è possibile estrarre, tramite la libreria PRAW, 1000 post per subreddit e 25 commenti per post appartenenti a una determinata categoria.

Le categorie da noi scelte sono:

- New: post più recenti
- Top: post con punteggio più alto

Per quanto riguarda le difficoltà incontrate per effettuare l'analisi del sentiment, non essendo ancora un campo stabile e "brevettato" al 100% , abbiamo speso molto tempo alla ricerca di librerie e pacchetti senza trovare l'efficienza desiderata. Poi, con TextBlob, siamo riusciti a coincidere affidabilità di analisi e compattezza del formato, adatti al nostro stile di programmazione, visto che anche l'estrazione dei dati è stata effettuata in linguaggio Python.

Possibili sviluppi e conclusioni

Al momento Reddit è il portale più seguito negli Stati Uniti per quanto riguarda le news, ed è molto utilizzato dai politici per le proprie campagne elettorali. Presenta perciò molte potenzialità di analisi e un domani potrebbe espandersi anche in Europa.

Il nostro intento è stato quello di stabilire chi, secondo gli utenti, è il candidato più popolare prendendo in considerazione due subreddit, ma è solo una delle tante possibili analisi da effettuare in questo campo.