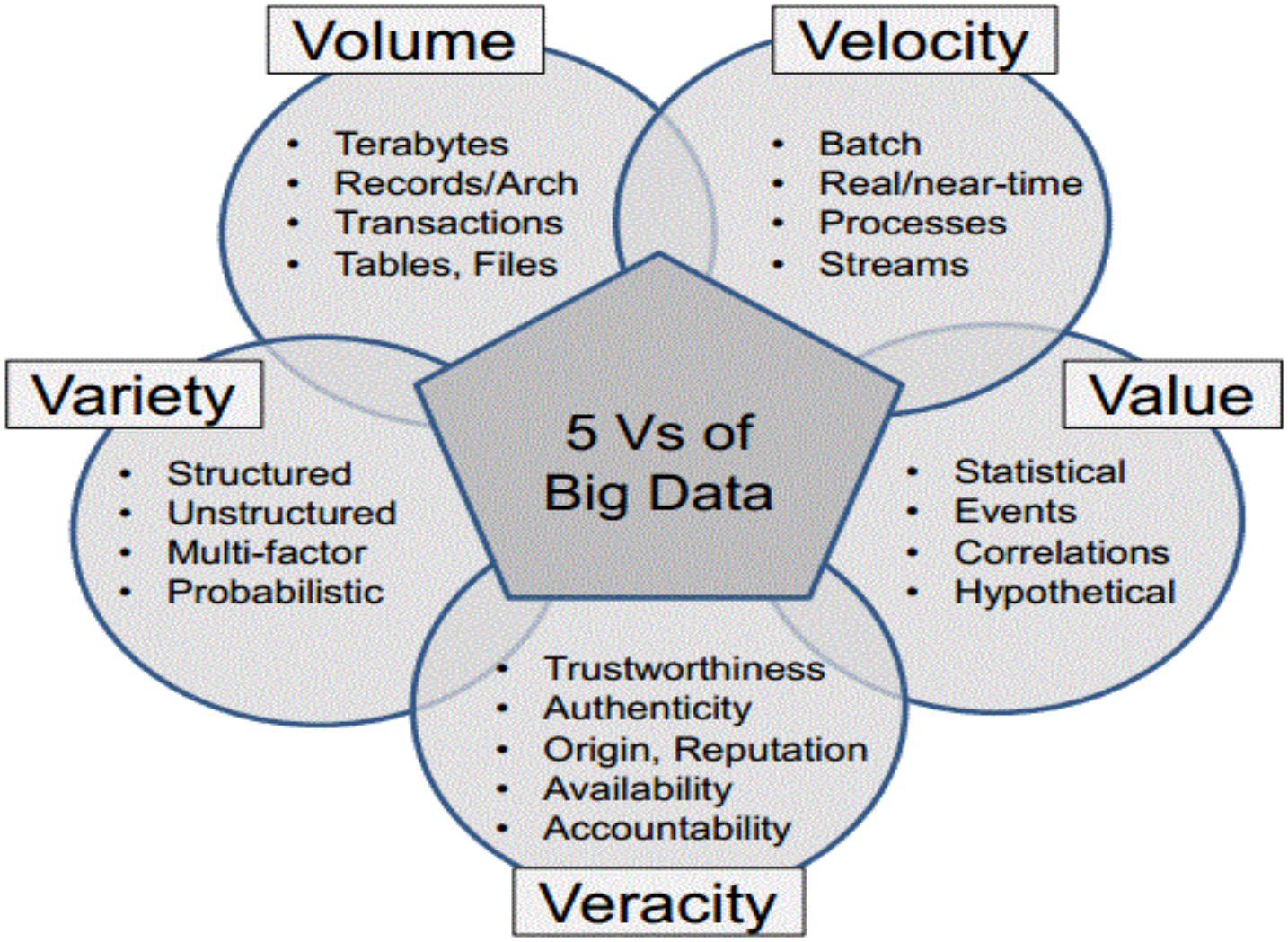
Cenni sui big data

Big Data

- Rappresentano uno dei fattori evolutivi nel mondo dell'analisi dei dati e della BI.
- La rete è una condizione necessaria per raccogliere e condividere i dati e grazie alle nuove tecnologie di raccolta, storage ed elaborazione e alla maggior capacità analitica e interpretativa di cui le imprese possono dotarsi, oggi le informazioni trattate si aggirano nell'ordine dei Petabytes fino a spingersi ai Zetabytes e i volumi sono in continuo aumento.
- Non esiste una definizione univoca di Big Data, con questo termine di indica un fenomeno nato a partire dagli anni Duemila che consiste nell'esplosione della quantità di informazioni disponibili.
- Nella letteratura Laney (2001) è il primo a soffermarsi, non nella descrizione dei metodi e scopi dei Big Data bensì, nella descrizione di 3 dimensioni che caratterizzano questo particolare tipo di dati.

Le 5 V dei Big Data



Secondo la TDWI (The Data Warehousing Institute), per riconoscere i Big Data è essenziale definire 3 requisiti base:

- Volume: indica la crescita esponenziale di informazione disponibile
- Velocità: è riferita alla velocità di raccolta e analisi dei dati
- Varietà: rappresenta l'eterogeneità dei dati a disposizione

1. Volume

Kilobyte (KB)	10^3
Megabyte (MB)	10^6
Gigabyte (GB)	10^9
Terabyte (TB)	10^{12}
Petabyte (PB)	10^{15}
Exabyte (EB)	10^{18}
Zetabyte (ZB)	10^{21}
Yottabite (YB)	10^{24}

Considera la dimensione in termini di Bytes dei database utilizzati per archiviare i dati aziendali. Ad oggi, però non è stata definita una soglia che distingua tra ciò che è Big Data e ciò che non lo è.

Dando uno sguardo alla tabella possiamo notare i multipli dei Bytes.

Secondo la TDWI (The Data Warehousing Institute), per riconoscere i Big Data è essenziale definire 3 requisiti base:

2. Velocity È la velocità con la quali i dati si generano, si raccolgono, si aggiornano e si elaborano. Riferita alla necessità di ridurre i tempi di gestione e analisi, poiché il dato può diventare obsoleto in pochissimo tempo.

3. Variety

Che può essere intesa come molteplicità di fonti o come eterogeneità di formato dei dati. In prima istanza, infatti si possono avere dati generati da diverse fonti interne o esterne. Questi dati poi possono avere diversi formati (*database, testo, video, immagini, audio eccetera*) riconducibili a tre categorie di dati; quelli strutturati, semi-strutturati e non strutturati

dati strutturati : es.

01 Paolo Verde

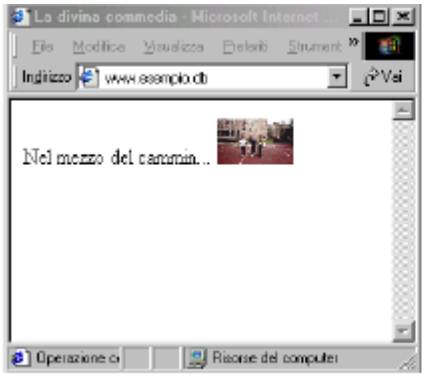
02 Mario Gialli

03 Grazia Bianchi

3. Variety

Che può essere intesa come molteplicità di fonti o come eterogeneità di formato dei dati. In prima istanza, infatti si possono avere dati generati da diverse fonti interne o esterne. Questi dati poi possono avere diversi formati (*database, testo, video, immagini, audio eccetera*) riconducibili a tre categorie di dati; quelli strutturati, semi-strutturati e non strutturati

dati semi strutturati




```
<html>
<title>La divina commedia</title>
Nel mezzo del cammin…
<img/></html>
```

DATI CON STRUTTURA PARZIALE

www.cs.unibo.it/~montesi/

3. Variety

Che può essere intesa come molteplicità di fonti o come eterogeneità di formato dei dati. In prima istanza, infatti si possono avere dati generati da diverse fonti interne o esterne. Questi dati poi possono avere diversi formati (*database, testo, video, immagini, audio eccetera*) riconducibili a tre categorie di dati; quelli strutturati, semi-strutturati e non strutturati

dati non strutturati es:

4. Value

Valore: riguarda le perdite di informazione dovute all’estrazione di conoscenza

5. Veracity

Veridicità: rappresenta l’inaffidabilità dei dati a disposizione

Quali sono le caratteristiche dei Big Data?

Le cosiddette «Nuove V» sono:

- Veridicità: rappresenta l'inaffidabilità dei dati a disposizione
- Valore: riguarda le perdite di informazione dovute all'estrazione di conoscenza

NOTE IN LETTERATURA « 5 V OF BIG DATA» (in figura)

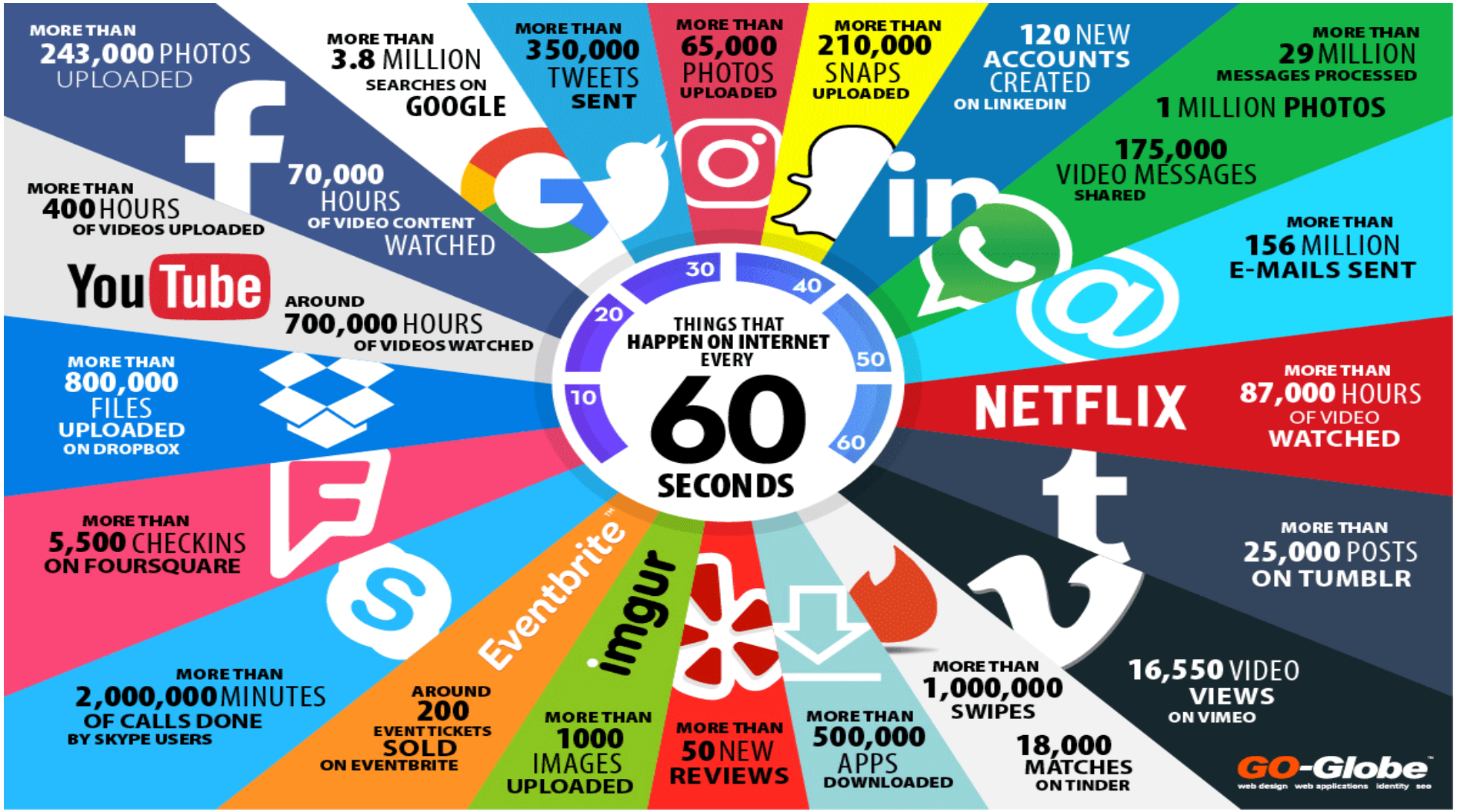
INSIEME ALLE PRIME 3 FORMANO LE COSÌ

- Variabilità: si riferisce alle sfumature di significato dovute al contesto in cui si verificano
- Visualizzazione: indica la rappresentazione delle informazioni chiave

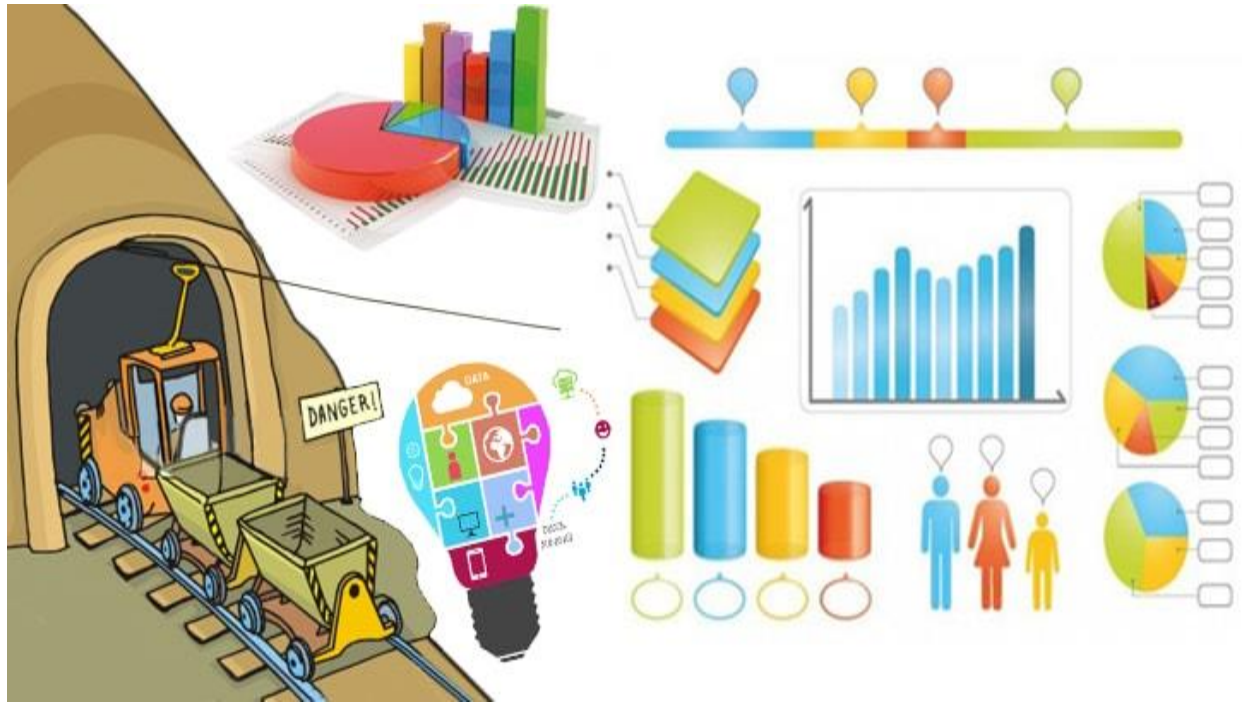
Privacy-by design

- È necessario un cambio nel concetto di proprietà del dato personale che ponga il singolo individuo al centro restituendogli trasparenza e diritti. In una parola democratizzare i Big Data.

- La grande sfida sta nel progettare ecosistemi per i dati personali che diano a tutti la possibilità di gestire le proprie informazioni personali e l'inter-scambio con le entità esterne, persone ed istituzioni, promuovendo trasparenza e trust.


MORE THAN 243,000 PHOTOS UPLOADED
MORE THAN 3.8 MILLION SEARCHES ON GOOGLE
MORE THAN 350,000 TWEETS SENT
MORE THAN 65,000 PHOTOS UPLOADED
MORE THAN 210,000 SNAPS UPLOADED
120 NEW ACCOUNTS CREATED ON LINKEDIN
MORE THAN 29 MILLION MESSAGES PROCESSED
1 MILLION PHOTOS
175,000 VIDEO MESSAGES SHARED
MORE THAN 400 HOURS OF VIDEOS UPLOADED
70,000 HOURS OF VIDEO CONTENT WATCHED
MORE THAN 156 MILLION E-MAILS SENT
AROUND 700,000 HOURS OF VIDEOS WATCHED
THINGS THAT HAPPEN ON INTERNET EVERY
60 SECONDS
MORE THAN 800,000 FILES UPLOADED ON DROPBOX
NETFLIX
MORE THAN 87,000 HOURS OF VIDEO WATCHED
MORE THAN 5,500 CHECKINS ON FOURSQUARE
MORE THAN 25,000 POSTS ON TUMBLR
MORE THAN 2,000,000 MINUTES OF CALLS DONE BY SKYPE USERS
AROUND 200 EVENT TICKETS SOLD ON EVENTBRITE
MORE THAN 1000 IMAGES UPLOADED
MORE THAN 50 NEW REVIEWS
MORE THAN 500,000 APPS DOWNLOADED
MORE THAN 1,000,000 SWIPES
18,000 MATCHES ON TINDER
16,550 VIDEO VIEWS ON VIMEO
GO-Globe
web design web applications identity seo

Text Mining: il processo di estrazione del testo



Il concetto fondamentale alla base del Text Mining è l'estrazione di poche informazioni rilevanti da un enorme ammontare di dati a disposizione.

«[…] *il potere risolutivo sta nelle parole significative*.» Hans Peter Luhn (1896-1964) Luhn definì che l'informazione statistica deriva dalla frequenza e dalla distribuzione delle parole e la utilizzò per calcolare una *misura* del significato di singole parole e poi di intere frasi.

Insieme al Data Mining nasce per superare i limiti delle tradizionali tecniche di analisi di un limitato numero di dati.

«*La scoperta da parte di un computer di nuove, in precedenza sconosciute, informazioni attraverso l'estrazione automatica dii differenti documenti scritti*» Hearst *(2003)*

Text Mining: il processo di estrazione del testo

Il processo si può riassumere in 3 fasi principali:

- Collezione dei dati
- Pre-processamento del testo
- Applicazione delle tecniche di Text Mining

Text Mining: pre-processo del testo

Pre-processamento del testo: adatta il testo grezzo in testo analizzabile operando una pre-elaborazione e pulizia dei dati per rimuovere le anomalie tramite:

- Tokenizzazione, rompe una sequenza di caratteri in parole o frasi chiamati token
- Filtraggio, consiste nella rimozione di parti del testo non ritenute rilevanti per l'analisi
- Lemmatizzazione, viene effettuata un'analisi morfologica delle parole raggruppando le varie forme flesse delle parole
- Derivazione, processo mediante il quale si crea un tema da una radice o parola pre-esistente

Text Mining: pre-processo del testo

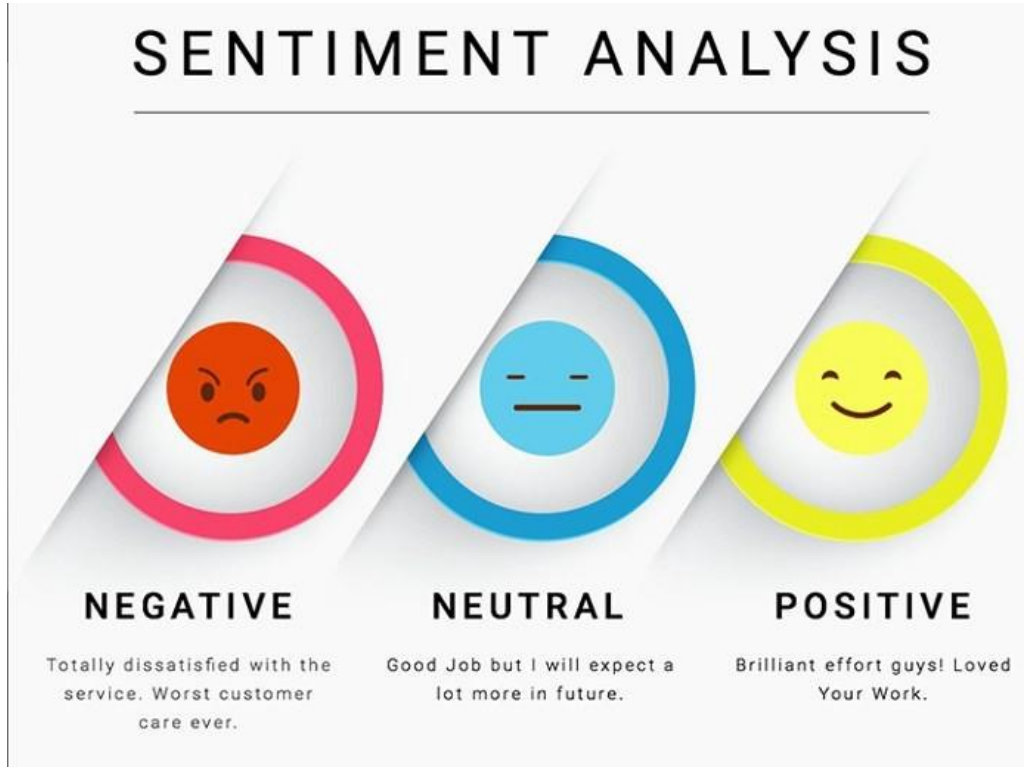
Applicazione delle tecniche di Text Mining: i dati testuali raccolti e adattati vengono analizzati tramite le tecniche di estrazione del testo.

Le tecniche di Text Mining hanno lo scopo di rintracciare le informazioni nascoste in un testo, sono strutturate secondo particolari algoritmi che selezionano le parti rilevanti di un testo.

Tra le più diffuse compaiono:

- Categorizzazione dei testi
- Estrazione dell'informazione
- Recupero delle informazioni
- Elaborazione del linguaggio naturale
- Clustering
- Riepilogo di testo
- Analisi del sentimento

Sentiment Analysis: analisi del sentimento



Questa tecnica è importante per estrarre informazioni soggettive dal contenuto utile a comprendere la risposta emotiva di un soggetto in un contesto.

La prima volta che comprare il termine Sentiment Analysis è in un testo del 2003 di *Nasasuka e Yi*.

Con gli anni diventa strumento indispensabile per conoscere la cosiddetta *brand perception* sfruttando lo scambio di interazioni degli utenti della rete.

Sentiment Analysis: analisi del sentimento

La Sentiment Analysis è strutturata principalmente su tre livelli:

- Documento: classifica l'intero documento indagando l'opinione contenuta in esso.
  IMPORTANTE: il documento deve contenere opinioni sulle singole entità
- Frase: viene indagata l'opinione contenuta in ciascuna frase soggettiva, bisogna quindi classificare le frasi in soggettive, oggettive e neutrali.
- Target: l'analisi dei sentimenti viene effettuata sulle entità e non necessariamente sull'intera frase.
  Es. «Anche se lo staff non era molto cordiale, l'hotel mi piace» le entità sono l'hotel e il servizio.

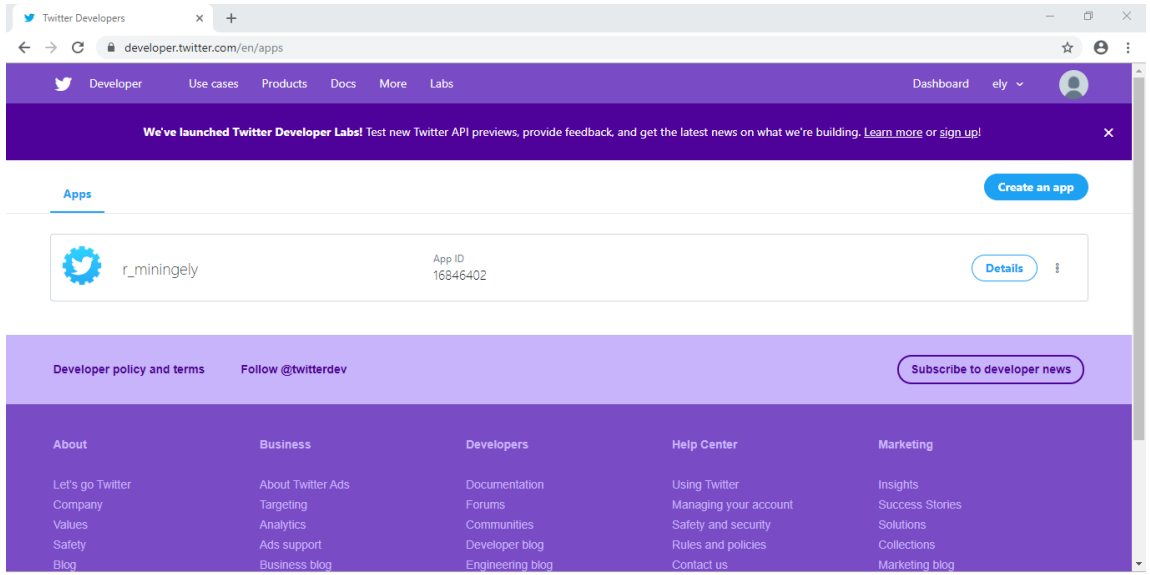
Sentiment Analysis: analisi del sentimento

- Si parla, anche in questo campo, di *SUPERAMENTO DELLE TRADIZIONALI TECNICHE STATISTICHE.* In quanto sondaggi e indagini statistiche vengono sostituiti da opinioni volontariamente rilasciate sulla rete.
- L'analisi è sviluppata a partire dall'uso di Opinion Words che vengono riconosciute dalla macchina e classificate sotto l'aspetto della positività o della negatività del sentimento che raffigurano.
- Problematiche: distinzione tra positivo o negativo, artefici linguistici, figure retoriche, emoticon, sarcasmo…

Cos’è Twitter?

- È una piattaforma gratuita di social network o in altre parole un servizio di micro-blogging
- È possibile condividere file digitali come foto o video e brevi messaggi, detti Tweets, con un massimo di 280 caratteri
- I post sono pubblici e leggibili da chiunque
  «Quello che dici su Twitter può essere visto istantaneamente in tutto il mondo»
- Ambiente ottimale per l’analisi sociale in quanto conta circa 220 milioni di utenti attivi
- Tramite il simbolo @ è possibile menzionare altri utenti
- È possibile creare discussioni su un particolare tema precedendo con # la parola che individua il topic richiesto
- È consentito condividere i pensieri di altri utenti tramite apposito comando, Retweet, o digitando le lettere RT all’inizio del post.

Le fasi dell'analisi: Registrazione al social network e creazione di API



La prima fase prevede:

- registrazione al social network
- creazione di un API (*Application Programming Interfaces*) che tramite protocollo *OAuth* rilascia un *Token* senza scadenza. Vengono generate le chiavi di accesso divise in due gruppi: Consumer API keys e Access token

Le fasi dell’analisi: Registrazione al social network e creazione di API

I pacchetti installati in questa fase sono “twitteR” e “RCurl”