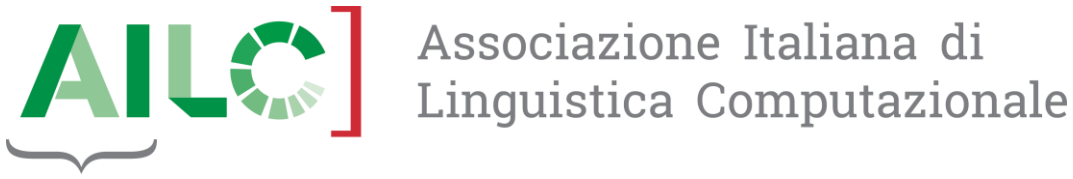
Trattamento Automatico delle Lingue Naturali e Data Science

Francesco Cutugno



Società di area linguistica presenti nell'assemblea delle consulte e delle associazioni di Area 10 del CUN

- Società Italiana di Glottologia
- Società di Linguistica Italiana
- Associazione Italiana di Linguistica Applicata
- Società Italiana di Didattica delle Lingue e Linguistica Educativa
- Associazione Italiana di Scienze della Voce
- Associazione Italiana di Linguistica Computazionale*

Il settore scientifico disciplinare di riferimento è L-Lin/01 e quello concorsuale è 10/G1, entrambe le declaratorie contengono riferimenti espliciti ai rapporti fra linguistica ed informatica, in un'ottica altamente interdisciplinare

*percorso di adesione in itinere

Associazioni cross-area: AISV e AILC

- Oltre che da L-Lin/01, ad AISV e AILC aderiscono soci provenienti da vari SSD quali ad es. Inf/01 (Area CUN 01), Ing-Inf/05 (Area CUN 09)

Fra i temi di interesse principali:

- Trattamento Automatico dei Testi e del Parlato
- Estrazione di conoscenza semantica
- Agenti conversazionali
- Traduzione automatica testo-testo e speech-speech
- Sentiment analysis e affective computing

L’80% dei Big Data sulla rete è costituita da testo non strutturato…

per non parlare di tutto l’audio che viene acquisito e trattato automaticamente dai sistemi conversazionali, youtube eccetera

Tech / #BigData

Forbes CommunityVoice™ Connecting expert communities to the Forbes audience. What is this?

JUN 5, 2017 @ 07:00 AM 7,536

The Big (Unstructured) Data Problem

Forbes Technology Council

Elite CIOs, CTOs & execs offer firsthand insights on tech & business. FULL BIO

Opinions expressed by Forbes Contributors are their own.

POST WRITTEN BY

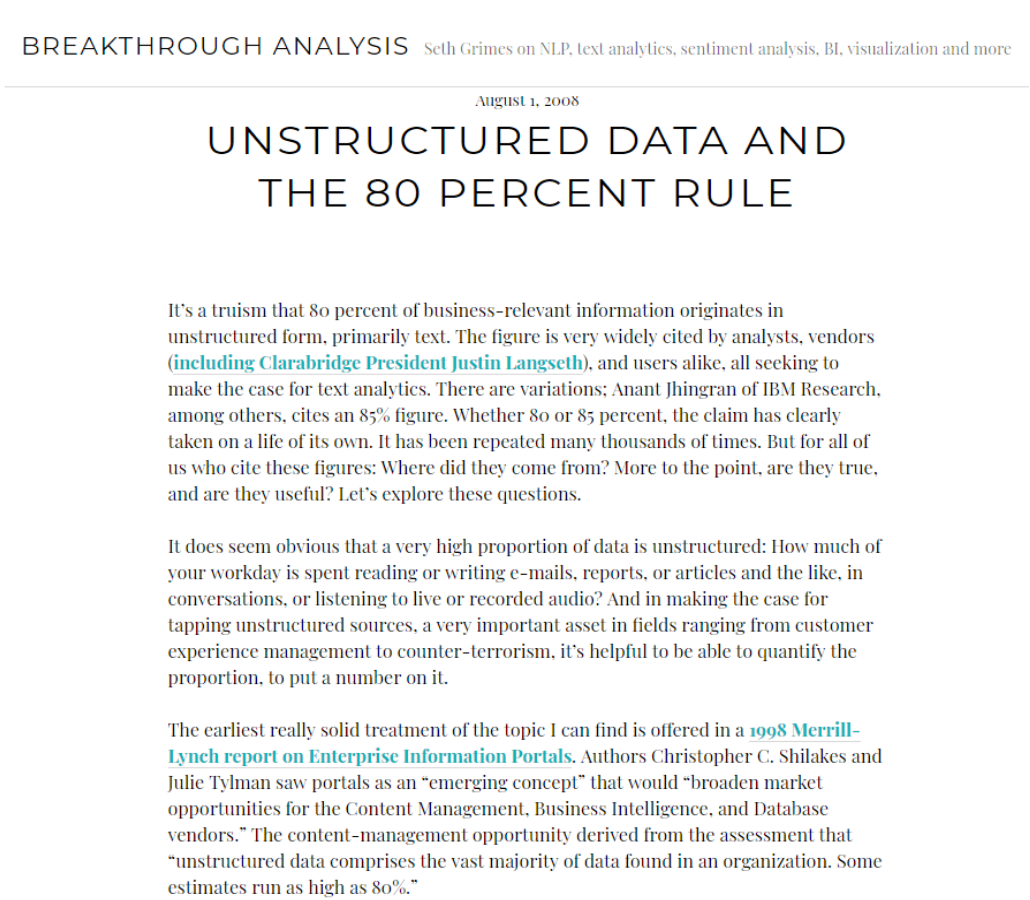
Juliette Rizkallah

Chief Marketing Officer at SailPoint, overseeing all aspects of the company marketing strategy, positioning and execution.

Juliette Rizkallah, Forbes Councils

The face of data breaches changed last year. The one that marked that change for me was the breach that involved former Secretary of State Colin Powell’s Gmail account. Targeted to expose the Hillary Clinton campaign, Colin Powell’s emails were posted on DCLinks.com for everyone to read. One of them had an attachment listing Salesforce’s acquisition targets and the details of its M&A strategy. Colin Powell, a member of Salesforce’s board, had access, through his personal email account, to sensitive information. When his personal email was hacked, all of that sensitive information was exposed -- and blasted out in the headlines.

https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem/#1324247b493a



BREAKTHROUGH ANALYSIS Seth Grimes on NLP, text analytics, sentiment analysis, BI, visualization and more

August 1, 2008

UNSTRUCTURED DATA AND THE 80 PERCENT RULE

It’s a truism that 80 percent of business-relevant information originates in unstructured form, primarily text. The figure is very widely cited by analysts, vendors (including Clarabridge President Justin Langseth), and users alike, all seeking to make the case for text analytics. There are variations; Anant Jhingran of IBM Research, among others, cites an 85% figure. Whether 80 or 85 percent, the claim has clearly taken on a life of its own. It has been repeated many thousands of times. But for all of us who cite these figures: Where did they come from? More to the point, are they true, and are they useful? Let’s explore these questions.

It does seem obvious that a very high proportion of data is unstructured: How much of your workday is spent reading or writing e-mails, reports, or articles and the like, in conversations, or listening to live or recorded audio? And in making the case for tapping unstructured sources, a very important asset in fields ranging from customer experience management to counter-terrorism, it’s helpful to be able to quantify the proportion, to put a number on it.

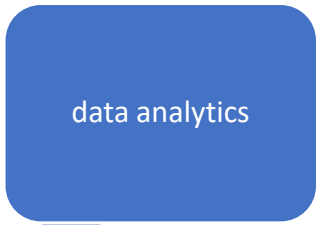
The earliest really solid treatment of the topic I can find is offered in a 1998 Merrill-Lynch report on Enterprise Information Portals. Authors Christopher C. Shilakes and Julie Tylman saw portals as an “emerging concept” that would “broaden market opportunities for the Content Management, Business Intelligence, and Database vendors.” The content-management opportunity derived from the assessment that “unstructured data comprises the vast majority of data found in an organization. Some estimates run as high as 80%.”

https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Il concetto di risorsa linguistica

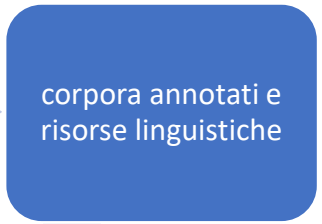
- In ambito Data Science gli operatori del Natural Language Processing (NLP) considerano il loro primario oggetto di studi il dato linguistico quantitativo, raccolto da ogni possibile fonte, vale a dire:
  - Testi scritti di qualsiasi natura
  - Parlato registrato digitalmente
  - Materiali multimodali (ad esempio voce+gesti)
  - Materiali multimediali (ad esempio filmati, ipertesti)
- la risorsa linguistica è il risultato della trattamento del dato linguistico arricchito di annotazioni di vario genere e fornito di metadatazione, attraverso processi elaborativi che richiedono competenze linguistiche
- Il lavoro di elaborazione finalizzato alla ricerca ed alla produzione è eseguito su risorse e non su dati linguistici grezzi

Una possibile pipeline

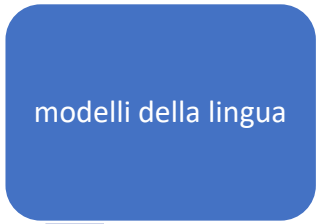


- Operazioni su materiali linguistici "grezzi"

- Lessico condiviso fra linguisti e informatici



- Di prevalente competenza linguistica



- Processo di astrazione e generalizzazione

- Semantica (understanding), esecuzione, back-end per le applicazioni

I tre quesiti

1) obiettivi culturali della classe di laurea, contenuti disciplinari e competenze trasversali

Obiettivi culturali:

- Riconoscere il ruolo cruciale delle risorse linguistiche all'interno delle Data Sciences
- Potenziamento della interdisciplinarità Informatica-Linguistica

Contenuti:

- di Information Extraction e Data Analytics da risorse linguistiche
- NLP per la multimodalità e la multimedialità
- Linguistica e Web Semantico
- Il ruolo della multilingualità nella progettazione dei sistemi informativi

Competenze:

- Linguistica Generale
- Linguistica Computazionale e Speech Processing
- Information Retrieval

2) sbocchi professionali sia nel mondo del lavoro che nel mondo della ricerca

Aziende Italiane:

Cedat85, CELI, Expert System, Euregio, Interactive Media, NTTData, Pervoice Almaviva, Questit, QuestionCube, Reveal, Spazio Dati, Spitch, Utopia.ai… e ancora altre

Multinazionali:

Nuance e Amazon a Torino, Google a Zurigo, Facebook a Parigi

Nell'accademia la ricerca nel settore NLP è già molto attiva in diversi centri italiani (Torino, Trento, Pisa, Bologna, Roma, Bari, Napoli, Cosenza), a forte indirizzo multidisciplinare, linguisti e informatici lavorano quasi sempre insieme sia in progetti di scuole dottorali che su temi di ricerca finanziata a livello nazionale e internazionale

3) quali elementi formativi devono essere aggiunti alla classe di laurea?

- Prerequisiti: crediti da recuperare in ambito informatico per laureati triennali provenienti da area umanistica e, similmente, crediti di linguistica generale da erogare agli studenti provenienti da area tecnica
- Stage presso le aziende precedentemente elencate
- Networking fra le istituzioni di ricerca già esistente ed in grado di supportare la circolazione e lo scambio di studenti e idee progettuali e di ricerca
- Focus group, scuole estive e attività di formazione extra-accademica con la collaborazione di tutte le associazioni scientifiche attive nel settore della linguistica in Italia.

…verso la creazione di una figura professionale e/o di ricerca orientata alla elaborazione di dati linguistici non- o semi- strutturati e multimodali…

… da spendere ad esempio:

- nel settore delle competenze di elaborazione dati multilingue,
- per la progettazione di agenti conversazionali,
- ovunque servano competenze da linguista in applicazioni ad alto impatto tecnologico