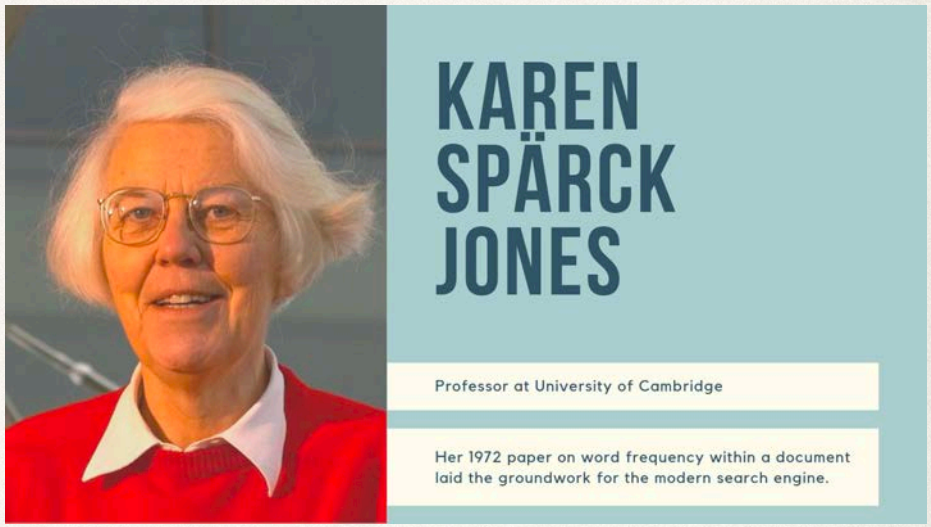
Karen Spärck Jones

Viviana Patti, Dipartimento di Informatica, University of Turin

viviana.patti@unito.it

Karen Spärck Jones (Regno Unito, 1935-2007)

- Percorsi non lineari e figure ibride
- Formazione umanistica: laurea in storia e filosofia
- Ambiente di ricerca:
  Centro di Ricerca sul linguaggio diretto da Margaret Masterman: filosofa e linguista, allieva di Wittgenstein, pioniera degli studi in linguistica computazionale, traduzione automatica
    - Wittgenstein's philosophy on language use

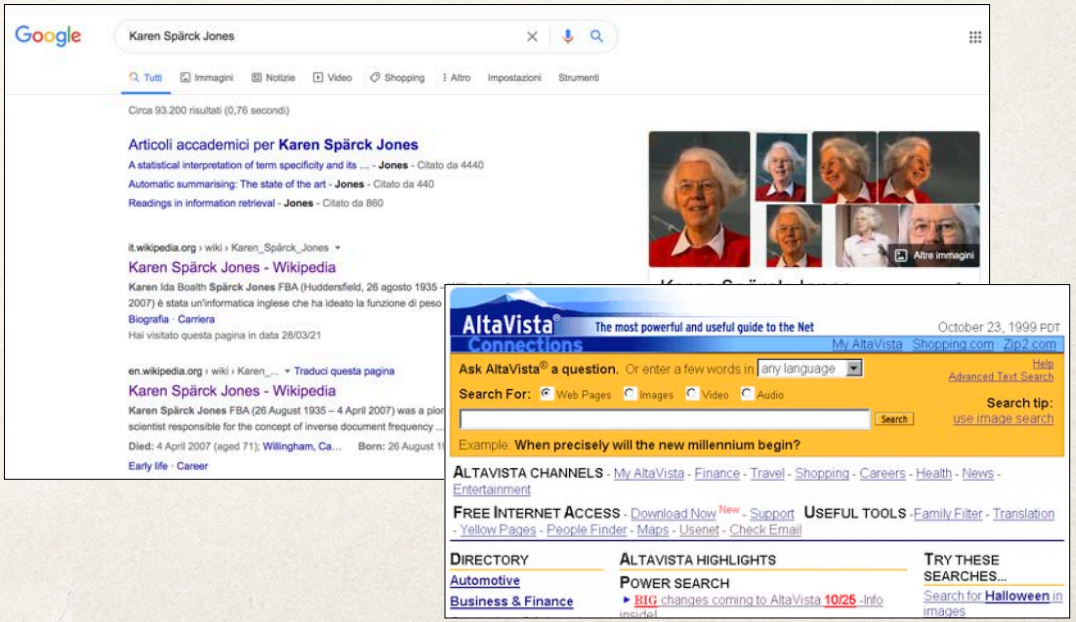


- 1964: Tesi di dottorato: "Sinonimi e classificazione semantica": applicazione di tecniche statistiche all'analisi del linguaggio, intuisce l'importanza delle risorse linguistiche, un elemento fondamentale per la linguistica computazionale nell'epoca dei big data

- 1972: contributo fondamentale: teoria statistica applicata al linguaggio
  concetto di TF-IDF: term frequency- inverse document frequency
    - Funzione di peso che vent'anni dopo sarà alla base del funzionamento dei primi motori di ricerca
- 1999: Full professor, Direttrice di Ricerca. Ricopre incarichi di prestigio e indirizzo in ACL. Premio a lei dedicato dalla comunità di Information Retrieval.
- Sposa Roger Needham ma mantiene il suo nome da nubile in ambito professionale:
  "E' comunque una cosa buona da fare sempre, mantiene la tua identità, se sai cosa significa"

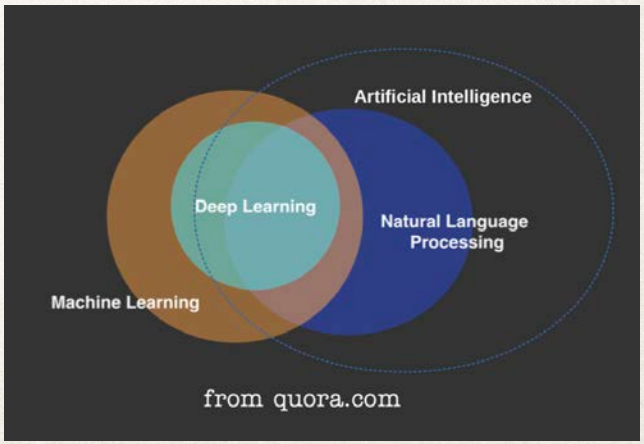
Karen Spärck Jones
Che cosa le dobbiamo

- Linguistica Computazionale
  - summarization, natural language interfaces, semantica lessicale, sistemi di dialogo…
- Traduttori automatici
- Motori di ricerca



Che cosa le dobbiamo
AI, NLP e CL

- Aree dell'Intelligenza Artificiale che si occupano specificamente del linguaggio umano si chiamano Linguistica Computazionale (CL) e Trattamento Automatico del Linguaggio (NLP Natural Language Processing)
- NLP include un'ampia varietà di compiti/task la cui soluzione si basa sulla comprensione del linguaggio umano.
- Esempi di task:
  - Machine translation
  - Part of Speech tagging (morphological analysis) and parsing (syntactic analysis)
  - Question answering
  - Sentiment analysis e opinion mining
  - Hate speech detection
  - Information retrieval
  - …



Che cosa le dobbiamo AI, NLP e CL

- Natural Language Processing (NLP) e Linguistica Computazionale (CL)
  - Morfologia → *how words are* Sintassi → *how words relate*
  - Semantica → *what words mean* Pragmatica → *what is the intention of the speaker* (contesto)
- Intuizione:
- Al pari di un essere umano, per poter comunicare in linguaggio umano un sistema di NLP deve
  - Acquisire la conoscenza relativa al linguaggio
  - Rappresentare al suo interno tale conoscenza
  - Utilizzare questa conoscenza in modo da poter affrontare in modo automatico i vari task:
- Ricevere domande e rispondere in modo pertinente, estrarre informazione per esempio sul sentiment o sull'odio, tradurre, cogliere l'intenzione ironica in un'espressione, ritrovamento di documenti rilevanti per l'utente nel mare magnum dei documenti disponibili sul web

Che cosa le dobbiamo
AI, NLP e CL

- I sistemi più recenti contengono poca conoscenza,
  ma sono in grado di apprenderla utilizzando algoritmi di apprendimento automatico supervisionato, da esempi con uno sforzo limitato
- Si costruiscono quindi delle grandi raccolte di esempi di uso del linguaggio, in cui
  - Regole, pattern ricorrenti e le irregolarità sono presenti e possono essere utilizzate dai sistemi per fare astrazione e generalizzare usando tecniche statistiche
- Deep Learning
- Profili multi-disciplinari fondamentali

Karen Spärck Jones
Che cosa le dobbiamo

- Linguistica Computazionale
- Motori di ricerca
  - 1995: AltaVista
    primo motore di ricerca
  - Trovare risultati in poco tempo
  - 1972: contributo fondamentale: teoria statistica applicata al linguaggio
    concetto di TF-IDF: term frequency- inverse document frequency

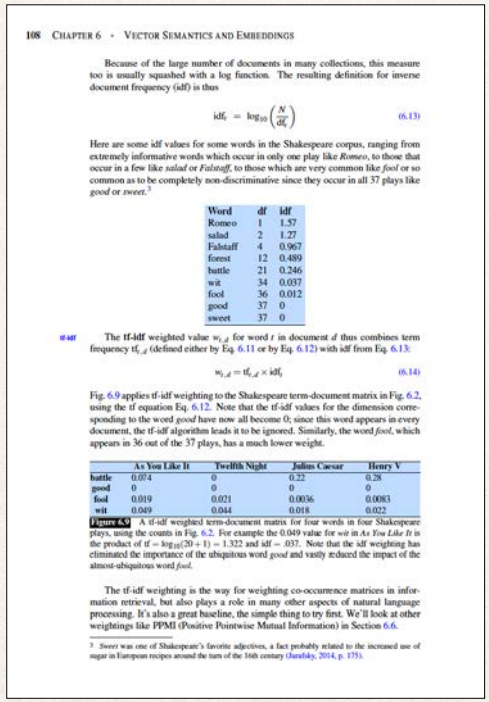
John I. Tait, 2007

well as in computing and artificial intelligence more broadly. In IR, her greatest contribution was probably her invention (1972) of the concept of inverse document frequency (IDF), in which the importance of terms is weighted according to the proportion of documents in the corpus in which they occur; the intuition being that terms which occur in many documents are poor index terms. This is the partial basis of all weighting schemes adopted by widely used Internet search engines, and will undoubtedly be a lasting contribution to the field.

  - Funzione di peso che vent'anni dopo sarà alla base del funzionamento dei primi motori di ricerca
  - Sguardo pionieristico: derivare informazione semantica su basi empiriche; importanza delle risorse linguistiche e di applicare la statistica all'analisi del linguaggio

Karen Spärck Jones
Che cosa le dobbiamo

- 1972: contributo fondamentale:
  teoria statistica applicata al linguaggio
  concetto di TF-IDF:
  term frequency- inverse document frequency



108 CHAPTER 6 • VECTOR SEMANTICS AND EMBEDDINGS

Because of the large number of documents in many collections, this measure too is usually squashed with a log function. The resulting definition for inverse document frequency (idf) is thus

$$\text{idf}_t = \log_{10}\left(\frac{N}{\text{df}_t}\right) \tag{6.13}$$

Here are some idf values for some words in the Shakespeare corpus, ranging from extremely informative words which occur in only one play like *Romeo*, to those that occur in a few like *salad* or *Falstaff*, to those which are very common like *fool* or so common as to be completely non-discriminative since they occur in all 37 plays like *good* or *sweet*.[3]

Word	df	idf
Romeo	1	1.57
salad	2	1.27
Falstaff	4	0.967
forest	12	0.489
battle	21	0.246
wit	34	0.037
fool	36	0.012
good	37	0
sweet	37	0

tf-idf

The **tf-idf** weighted value $w_{t,d}$ for word t in document d thus combines term frequency $\text{tf}_{t,d}$ (defined either by Eq. 6.11 or by Eq. 6.12) with idf from Eq. 6.13:

$$w_{t,d} = \text{tf}_{t,d} \times \text{idf}_t \tag{6.14}$$

Fig. 6.9 applies tf-idf weighting to the Shakespeare term-document matrix in Fig. 6.2, using the tf equation Eq. 6.12. Note that the tf-idf values for the dimension corresponding to the word *good* have now all become 0; since this word appears in every document, the tf-idf algorithm leads it to be ignored. Similarly, the word *fool*, which appears in 36 out of the 37 plays, has a much lower weight.

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	0.074	0	0.22	0.28
good	0	0	0	0
fool	0.019	0.021	0.0036	0.0083
wit	0.049	0.044	0.018	0.022

Figure 6.9 A tf-idf weighted term-document matrix for four words in four Shakespeare plays, using the counts in Fig. 6.2. For example the 0.049 value for *wit* in *As You Like It* is the product of $\text{tf} = \log_{10}(20 + 1) = 1.322$ and $\text{idf} = .037$. Note that the idf weighting has eliminated the importance of the ubiquitous word *good* and vastly reduced the impact of the almost-ubiquitous word *fool*.

The tf-idf weighting is the way for weighting co-occurrence matrices in information retrieval, but also plays a role in many other aspects of natural language processing. It's also a great baseline, the simple thing to try first. We'll look at other weightings like PPMI (Positive Pointwise Mutual Information) in Section 6.6.

[3] *Sweet* was one of Shakespeare's favorite adjectives, a fact probably related to the increased use of sugar in European recipes around the turn of the 16th century (Jurafsky, 2014, p. 175).

A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL

KAREN SPARCK JONES
University of Cambridge Computer Laboratory

The exhaustivity of document descriptions and the specificity of index terms are usually regarded as independent. It is suggested that specificity should be interpreted statistically, as a function of term use rather than of term meaning. The effects on retrieval of variations in term specificity are examined, experiments with three test collections showing in particular that frequently-occurring terms are required for good overall performance. It is argued that terms should be weighted according to collection frequency, so that matches on less frequent, more specific, terms are of greater value than matches on frequent terms. Results for the test collections show that considerable improvements in performance are obtained with this very simple procedure.

EXHAUSTIVITY AND SPECIFICITY

WE ARE FAMILIAR with the notions of exhaustivity and specificity: exhaustivity is a property of index descriptions, and specificity one of index terms. They are most clearly illustrated by a simple keyword or descriptor system. In this case the exhaustivity of a document description is the coverage of its various topics given by the terms assigned to it; and the specificity of an individual term is the level of detail at which a given concept is represented.

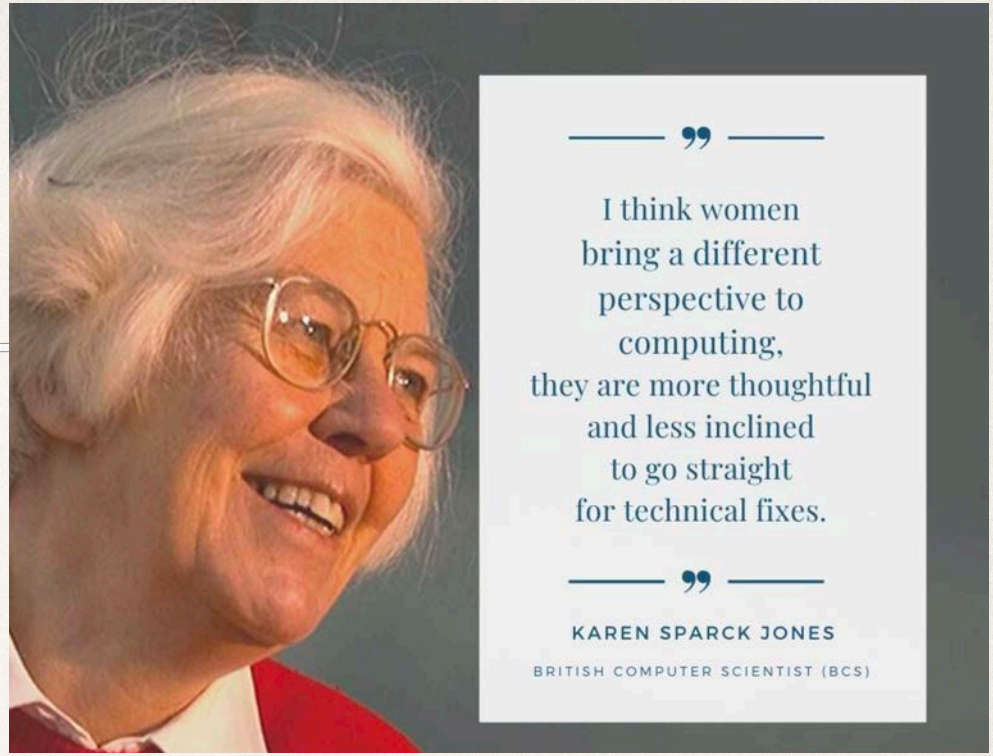
These features of a document retrieval system have been discussed by Cleverdon *et al.*[4] and Lancaster,[6] for example, and the effects of variation in either have been noted. For instance, if the exhaustivity of a document description is increased by the assignment of more terms, when the number of terms in the indexing vocabulary is constant, the chance of the document matching a request is increased. The idea of an optimum level of indexing exhaustivity for a given document collection then follows: the average number of descriptors per document should be adjusted so that, hopefully, the chances of requests matching relevant documents are maximized, while too many false drops are avoided. Exhaustivity obviously applies to requests too, and one function of a search strategy is to vary request exhaustivity. I shall be mainly concerned here, however, with document descriptions.

Specificity as characterized above is a semantic property of index terms: a term is more or less specific as its meaning is more or less detailed and precise. This is a natural view for anyone concerned with the construction of

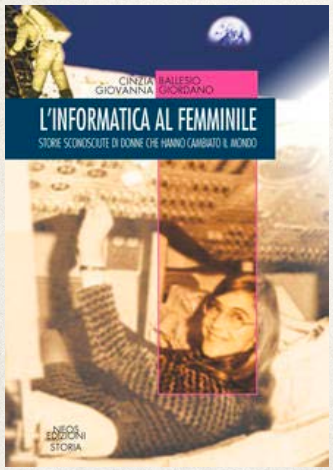
More women in tech

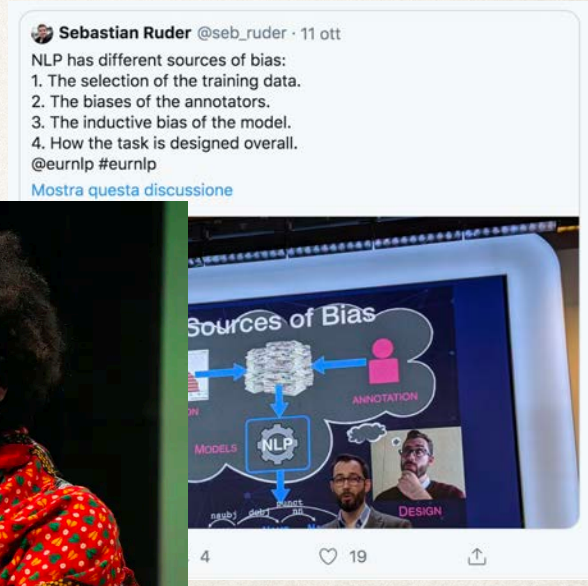
- Nell'era digitale è imprescindibile: more women working in tech
- Immaginario collettivo: informatica è maschile
  - stereotipo del nerd, giovani maschi soli in una stanza chini su una tastiera
- Impatto sociale
  - Technology is pervasive and women must be active part of this digital revolution, also to be able to recognise, and counteract the misogynistic behaviours we are observing online
  - Agenti conversazionali: lo strano caso del chatbot Tay

More women in tech



- Rischio:
  - Introdurre bias nei corpora e nei modelli addestrati sui dati
- Inclusive design & diversity
- "Credo anche che l'informatica sia intellettualmente affascinante, soprattutto se si ha intenzione di creare qualcosa che ancora non esiste"





Riferimenti

- ACL Lifetime Achievement Award:
  Karen Spärck Jones (2005) Some Points in a Time
  https://www.aclweb.org/anthology/J05-1001.pdf

- Karen Spärck Jones (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28:11–21
  https://doi.org/10.1108/00220410410560573

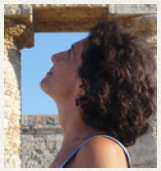
- Karen Spärck Jones (March 2007) Natural Language and the Information Layer: *Professor Karen Spärck Jones' health did not permit her to receive the* ACM Athena Award (at SIGIR 2007, 23–27 July 2007, Amsterdam) and BCS Lovelace Medal *in person, so she prepared this video recorded talk instead*. Length 33 minutes, 16:9 format. Available here:
  https://www.cl.cam.ac.uk/misc/obituaries/sparck-jones/video/

- John Tait (2007). Obituary. Karen Spärck Jones. Comput. Linguistics 33(3): 289-291
  https://www.aclweb.org/anthology/J07-3001.pdf

- Daniel Jurafsky & James H. Martin (2020) Speech and Language Processing. Draft of December 30, 2020. Chapter 6.
  https://web.stanford.edu/~jurafsky/slp3/6.pdf

- Cinzia Ballesio, Giovanna Giordano (2019) L'informatica al femminile. Storie sconosciute di donne che hanno cambiato il mondo
  https://informaticaalfemminile.it/

Who am I?



- Msc Computer Science in Philosophy
  - Multidisciplinary curriculum (analysis of language and computer science): philosophy of language, computational linguistics, cognitive science, programming, artificial intelligence.
- PhD in Computer Science
- Associate Professor at Computer Science Department, University of Turin

di.unito.it

- co-founding member of the Logic, Language and Cognition center, University of Turin (Philosophy, Computer Science, Psychology)
- Past member of Guarantee Committee at UniTo

- https://www.unito.it/persone/vpatti

The multi-disciplinary team

Simona Frenda

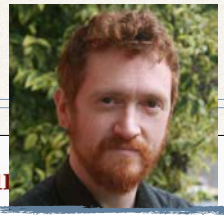
Komal Florio

Sohail Aktar

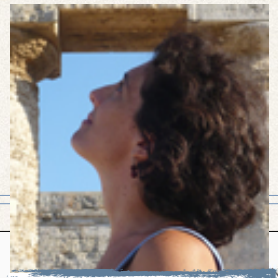
Elisa Di Nuovo

Endang W. Pamungkas

Alessandra T. Cignarella



Valerio BASILE

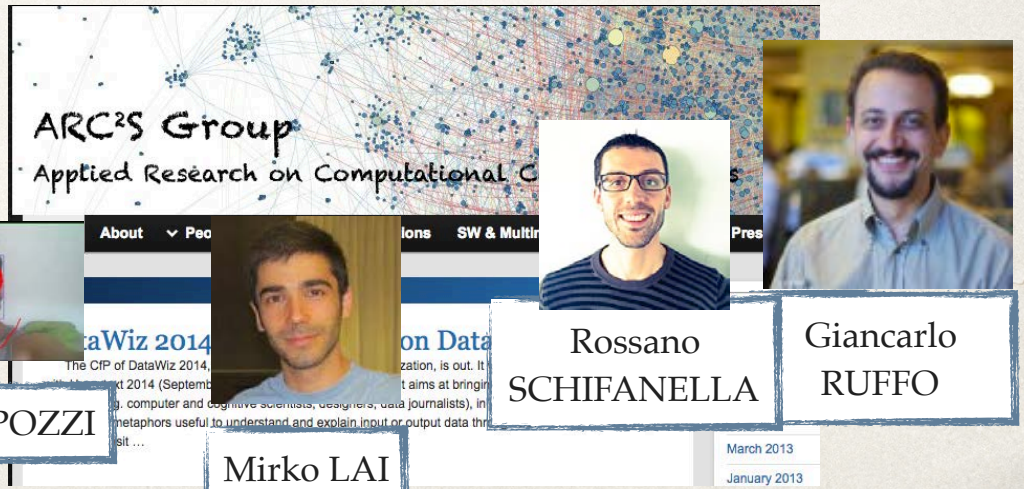


Viviana PATTI

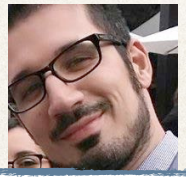
Cristina BOSCO

Manuela SANGUINETTI

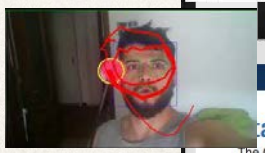
Delia Irazu HERNANDEZ FARIAS



Fabio POLETTO



Marco STRANISCI



Arthur CAPOZZI

Mirko LAI

Rossano SCHIFANELLA

Giancarlo RUFFO

Thank you!

viviana.patti@unito.it
https://www.unito.it/persone/vpatti