Text mining e testi biomedici. Una ricerca sulla continuità dell'informazione

Stefano Ballerio

Indicatori della continuità assistenziale
Melegnano, 25 maggio 2009

Agenda

- Il sovraccarico informativo e il text mining in campo biomedico
- La ricerca: lettere di dimissione e text mining
- Il text mining: fasi, metodi, strumenti
- Bibliografia

Il problema: il sovraccarico informativo e la necessità di conoscenze

«The volume of published **biomedical research**, and therefore the underlying biomedical **knowledge base**, is **expanding** at an increasing rate». Cohen e Hersh (2005)

PubMed:

- abstract di oltre 17.000.000 di *research papers*
- 40.000 nuovi abstract al mese

Da queste basi di dati è necessario **estrarre conoscenze** in modo efficiente.

Una soluzione: il text mining e la knowledge extraction

«Among the tools that can aid researchers in coping with this information overload are **text mining** and **knowledge extraction**».
Cohen e Hersh (2005)

Text mining

«**Text mining** can be broadly defined as a knowledge-intensive process in which a **user** interacts with a **document collection** over time by using a suite of **analysis tools**. In a manner analogous to data mining, text mining seeks to extract useful information from data sources through the **identification and exploration of interesting patterns**. In the case of text mining, however, tha data sources are document collections, and interesting patterns are found not among formalized database records but in the unstructured textual data in the documents in these collections». Feldman e Sanger (2007)

Applicazioni (1/2)

- Corporate finance e business intelligence

  tendenze e associazioni ricorrenti in relazione a transazioni, compagnie, prodotti, persone

- In campo biomedico, classificazione della letteratura scientifica

  costruzione di database annotati di articoli, abstract, relazioni ecc.

- Hypothesis generation in complementary structures in disjoint literatures

  TM su n articoli: A influisce su B

  TM su m articoli: B influisce su C

  allora (ipotesi) A influisce su C

  Swanson (1991) correla così emicrania e carenza di magnesio.

Applicazioni (2/2)

- Database di referti radiologici + TM = conoscenze epidemiologiche sui pazienti sottoposti a esame radiografico

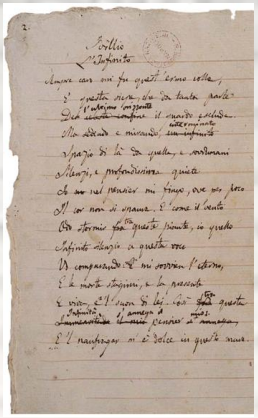
  Fiszman, Chapman, Aronsky, Evans e Haug (2000)

  Hripcsak, Austin, Alderson e Friedman (2002)

- Lettere di dimissione + TM = rilevamento delle complicanze

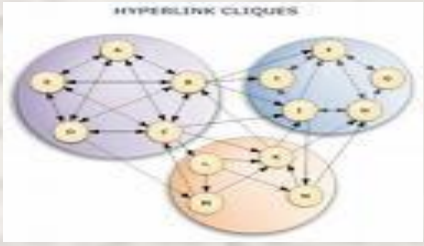
  Melton e Hripcsak (2005) applicano TM a un database di lettere di dimissione e rilevano le complicanze definite dal New York Patient Occurrence Reporting and Tracking System.

Fasi del text mining



➢text preprocessing (NLP, IE)

➢costituzione della collezione (IR)



➢obiettivi della ricerca
➢euristica: concetti e relazioni
➢traduzione dell'euristica in termini linguistici e statistici

➢mining (DM)

➢interpretazione e valutazione dei risultati

IR, NLP, IE

TM integra diverse tecniche per il trattamento del linguaggio e dell'informazione:

- **information retrieval** (IR)
- **natural language processing** (NLP)
- **information extraction** (IE).

L'obiettivo è individuare i materiali testuali necessari, processarli automaticamente ed estrarne informazioni.

Data mining

Analisi dei dati strutturati, di tipo quantitativo e numerico ➡ tecniche di **data mining** (DM)

DM integra metodi statistici e informatici per estrarre nuove conoscenze da grandi basi di **dati strutturati**.

La ricerca sulle lettere di dimissione

Obiettivi della ricerca:

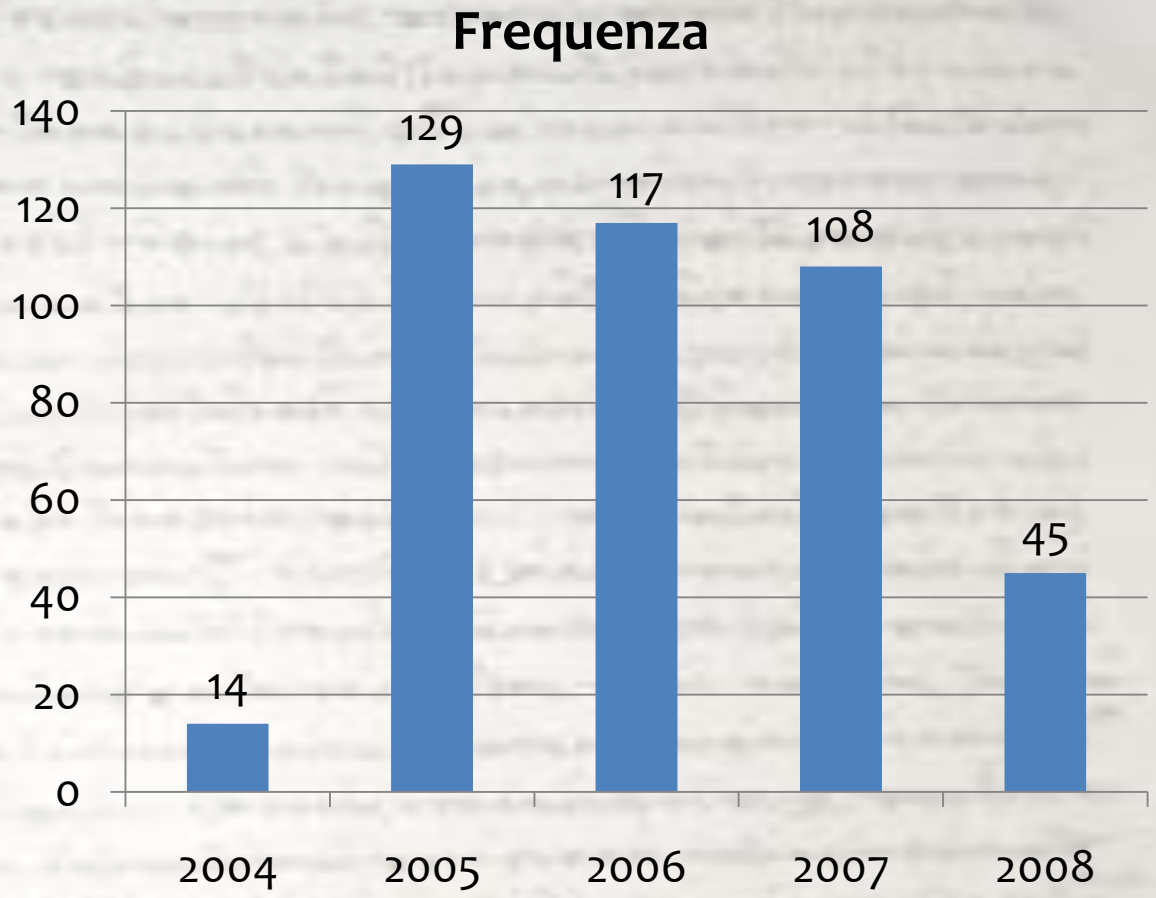
- valutare la **continuità assistenziale** limitatamente alla sua **dimensione informazionale**
- sperimentare **indicatori** di continuità assistenziale
- sviluppare e applicare **strumenti di analisi automatizzata dei testi**.

Il modello della Joint Commission

- **Accesso e continuità dell'assistenza (ACC), standard 3.2**: la cartella clinica contiene una copia della lettera di dimissione.
- La lettera di dimissione contiene:
  - il motivo del ricovero
  - riscontri e accertamenti fisici e di altro genere significativi
  - diagnosi e comorbilità significative
  - procedure diagnostiche e terapeutiche eseguite
  - terapia farmacologica significativa e altre terapie significative
  - condizioni del paziente alla dimissione
  - terapia farmacologica alla dimissione, tutti i farmaci da assumere al domicilio
  - **istruzioni di follow-up**.

I testi (1/2)

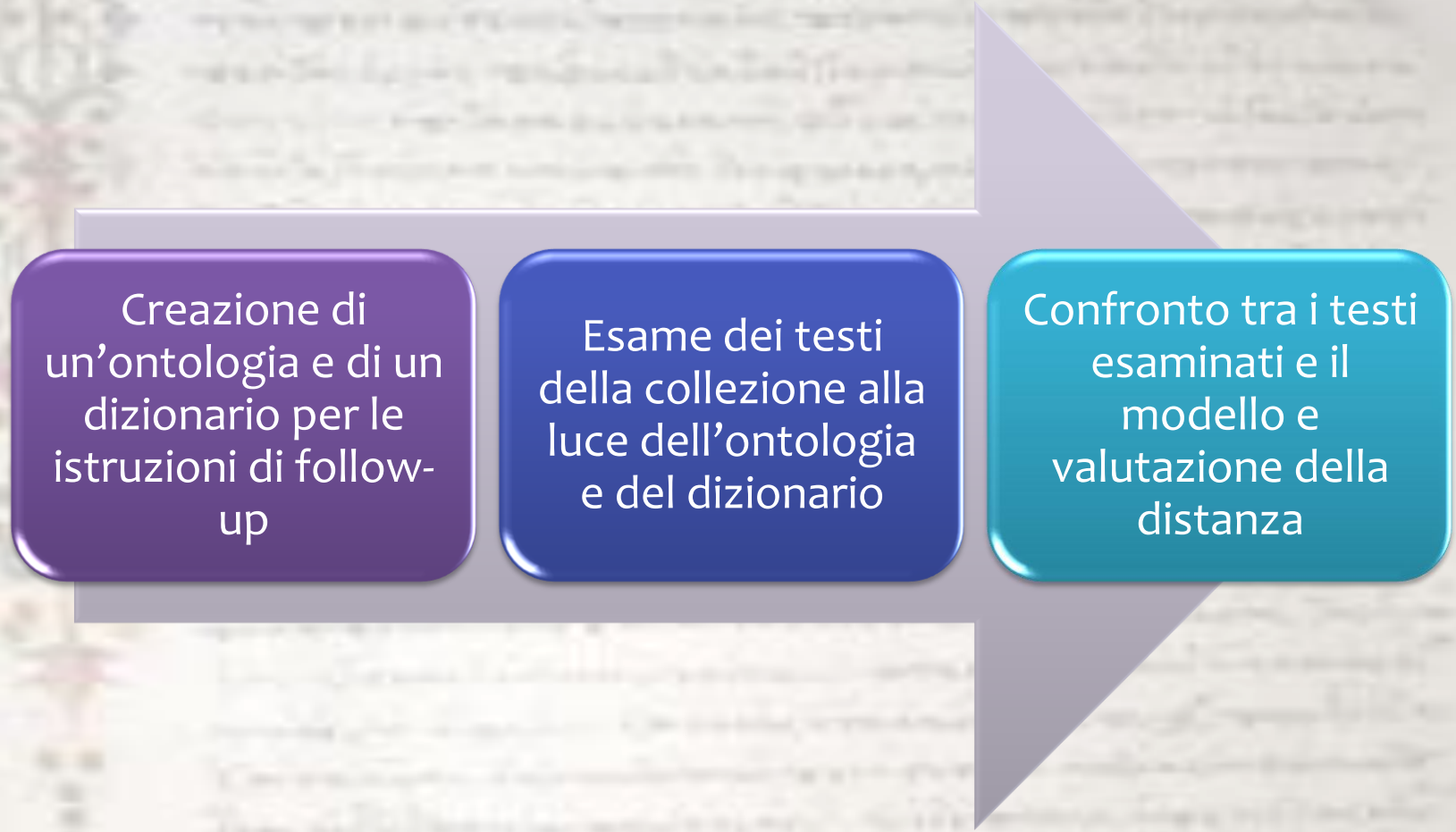
413 lettere di dimissione del reparto di cardiologia dell'ospedale Uboldo di Cernusco sul Naviglio. Le lettere sono relative a pazienti con scompenso cardiaco. L'intervallo temporale va dal 2004 al 2008.



I testi (2/2)

- I casi di scompenso cardiaco sono stati selezionati dal file delle **SDO** utilizzando i criteri **AHRQ** di ospedalizzazione evitabile per **scompenso cardiaco**.
- Dal file delle SDO si è risaliti alle **lettere di dimissione** tramite il numero di cartella clinica.
- Le lettere di dimissione sono in **formato Access™** e contengono quattro campi memo, relativi ad anamnesi, esami effettuati, decorso e prescrizioni alla dimissione.

Il confronto tra i testi e il modello



Un'ontologia e un dizionario per le istruzioni di follow-up

Istruzioni per il follow-up: i concetti

La composizione del dizionario: stop word e simboli eliminati

a
a causa
a fronte
a questo proposito
ad
agli
ai
al
all'
alla
alle
allo
altresì
anche
ancora
che

ci
circa
come
comunque
con
così…

1
2
3
4
5…

mg
cpr…

1. Furosemide 25 mg 1 cpr. Captopril 25 mg x 3. Allopurinolo 100 mg. Carvedilolo 6,25 mg 1/2 cpr x 2. Spironolattone 25 mg 1 cpr. Acido Acetilsalicilico 100 mg. Dieta ipocalorica (1600 Cal) Si raccomanda astensione dal fumo.

2. Furosemide. Captopril. Allopurinolo. Carvedilolo. Spironolattone. Acido Acetilsalicilico. Dieta ipocalorica Cal raccomanda astensione fumo.

La composizione del dizionario: tokenization e frequenze

WORD	COUNT	PERCENT
FUROSEMIDE	409	1,492973
CONTROLLO	403	1,471071
COMPENSO	355	1,295857
INR	254	0,927176
TERAPIA	223	0,814017
NON	220	0,803066
QUADRO	218	0,795766
MIGLIORAMENTO	202	0,737361
RENALE	168	0,613251
FUNZIONE	162	0,591349
BISOPROLOLO	153	0,558496
SCOMPENSO	141	0,514692
LASIX	139	0,507392
CONTROLLI	138	0,503742

La composizione del dizionario: lessico delle istruzioni di follow-up

WORD	COUNT	PERCENT
CONTROLLO	403	1,471071
CONTROLLI	138	0,503742
RACCOMANDA	98	0,35773
REGIME	69	0,251871
PESO	67	0,24457
MONITORAGGIO	60	0,219018
CONSIGLIA	59	0,215368
PONDERALE	41	0,149662
RACCOMANDANO	28	0,102208
RISPARMIO	28	0,102208
ESEGUIRE	24	0,087607
SEGNALARE	24	0,087607
CORPOREO	22	0,080307
RIPOSO	22	0,080307

Ontologia e dizionario: concetti e termini relativi

controllare peso

monitoraggio peso

verifica peso corporeo

controllo peso

verificare peso corporeo

osservazione peso

monitorare peso

Controllo del peso

osservazione peso corporeo

verifica peso

controllo peso corporeo

autovalutazione peso corporeo

verificare peso

controllare peso corporeo

autovalutazione peso

monitorare peso corporeo

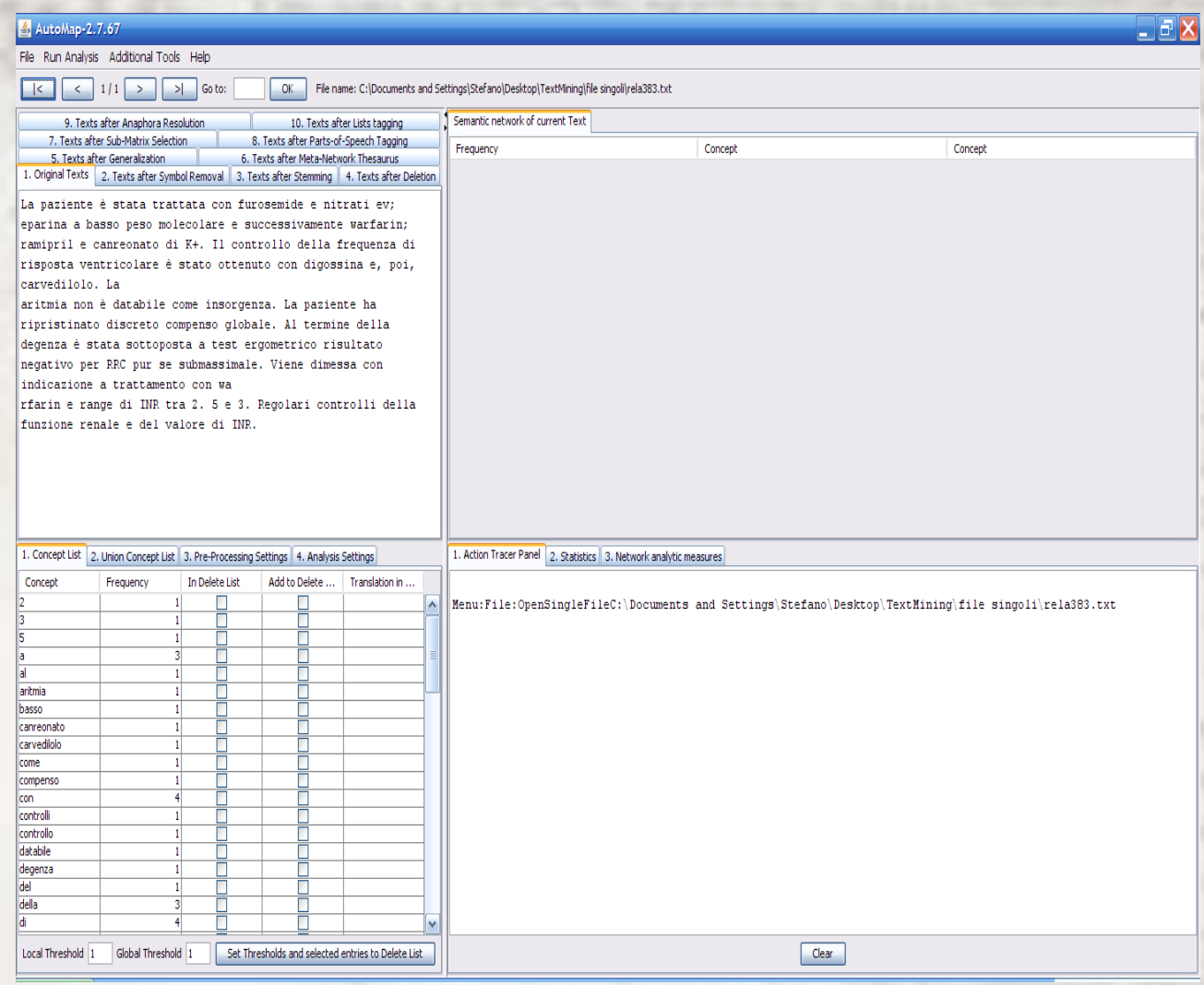
monitoraggio peso corporeo

L'esame della collezione: algoritmi per la ricerca

- **if** (controllo peso) or
  (controllare peso) or
  (monitoraggio peso) or
  (monitorare peso) or
  (autovalutazione peso) or
  (osservazione peso) or
  (verifica peso) or
  (verificare peso) or
  …
  (verificare peso corporeo)
  then peso, controllo
- **else** ¬ peso, controllo

Si ripete su tutti i testi della collezione, per tutti i concetti

L'esame della collezione: CASOS AutoMap



Per esaminare la collezione, abbiamo usato **AutoMap** di Kathleen M. Carley (CASOS – Carnegie Mellon University)

L’esame della collezione: risultati

Concetto	Occorrenze	Testi positivi	Percentuale
Attività	1	1	0,24
Riposo	22	22	5,33
Dieta	24	22	5,33
Perdita di peso	39	37	8,96
Controllo del peso	12	11	2,66
Alcol	3	3	0,73
Fumo	7	7	1,69
Altre categorie	0	0	0

Validazione

- Un **lettore umano** esamina i testi della collezione e verifica le risposte del sistema.
- Risultati
  - **sensibilità** del sistema: **0,484**

    nel dizionario manca *vita di risparmio*
  - **specificità** del sistema: **0,834**

    il sistema non discrimina *si è registrato un forte calo ponderale*

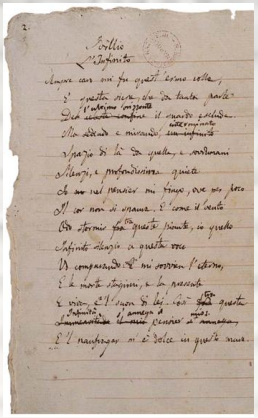
Analisi dei risultati del sistema

- Ampliare le **basi di conoscenze**:
  - per aumentare la sensibilità, è necessario ampliare il dizionario mediante l'esame di una collezione di testi più numerosa.
- Sviluppare le funzioni di **natural language processing**:
  - per aumentare la specificità, occorre un software con funzioni di NLP più raffinate.

Sviluppi

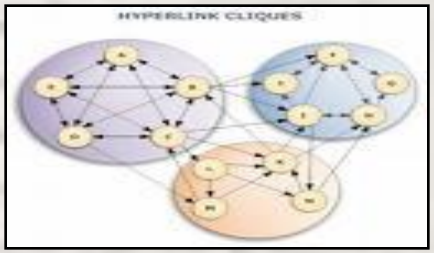
- Stiamo provando **NooJ** di Max Silberztein, un software specifico per il natural language processing.
- Abbiamo preparato un dizionario di circa 12.300 specialità e principi farmacologici, che NooJ può applicare su nuove collezioni di testi.

Fasi del text mining



➢text preprocessing (NLP, IE)

➢costituzione della collezione (IR)



➢obiettivi della ricerca
➢euristica: concetti e relazioni
➢traduzione dell'euristica in termini linguistici e statistici

➢mining (DM)

➢interpretazione e valutazione dei risultati

Il text preprocessing: obiettivo

- Obiettivo del text preprocessing è **trasformare ogni testo in una rappresentazione esplicitamente strutturata**.
- Il preprocessing è molto più complesso nel TM che nel DM: un testo è un oggetto strutturato (morfologia, sintassi, punteggiatura, testualità, redazione, grafica), ma lo è in modo complesso e non evidente.

Tratti: caratteri e parole

Quali **tratti** dei testi sono **rilevanti** per la loro rappresentazione?

Dipende dagli obiettivi e dall'euristica definiti.

Da un punto di vista linguistico, si va dai caratteri, alle parole, ai termini, ai concetti:

- **caratteri**
  - lettere, numeri e altri caratteri speciali
  - si possono cercare o eliminare (!, *, >)
- **parole**
  - ogni parola singola
  - si possono cercare solo alcune parole (entry list) o se ne possono eliminare (stop list: tipicamente, le parole funzione come gli articoli)

Tratti: termini e concetti

- **termini**
  - parole singole (*miocardio*) o espressioni composte (*Casa Bianca*, *struttura ospedaliera*)
  - possono essere ricondotti a termini normalizzati
- **concetti**
  - a uno stesso concetto (*astensione dall'alcol*) si riconducono diverse espressioni
  - il concetto include non solo varianti ortografiche (*astensione dall'alcol* o *astensione dall'alcool*) e morfologiche (*astenersi dall'alcol*), ma anche sinonimi (*astenersi da bevande alcoliche*) ecc.

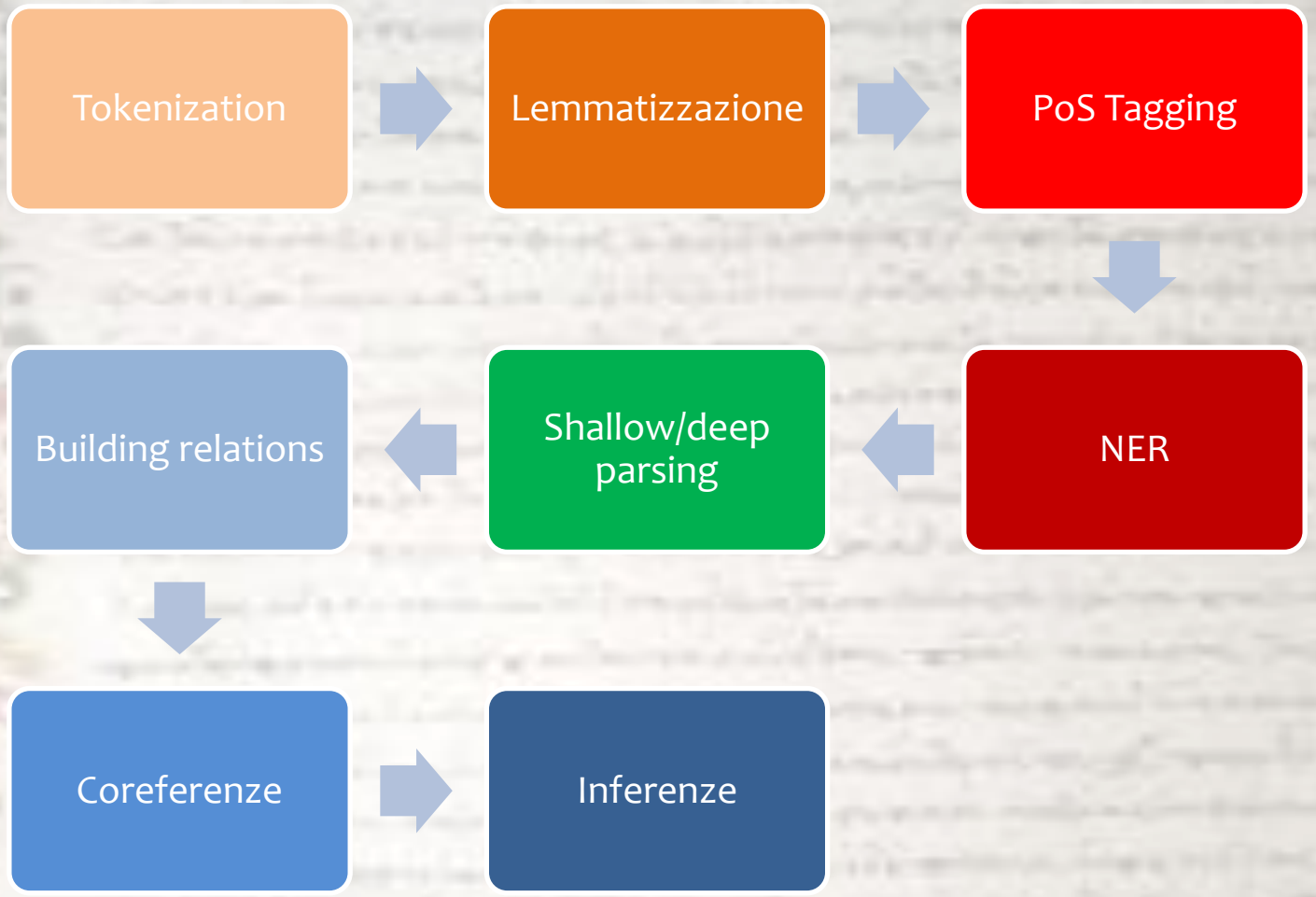
Tratti: come rappresentarli

Come rappresentare
la presenza/assenza
di un tratto?

a) con un'opposizione binaria 0/1

b) con una pesatura, sulla base dei valori di term frequency e inverse document frequency

Information flow del text preprocessing (1/4)



Information flow del text preprocessing (2/4)

1. **Tokenization**
   1. si scompone il testo in unità o **token**, ovvero in parole e poi in periodi e capoversi
   2. in questa fase si possono eliminare le **stop word**
   3. esempio: *~~il~~ paziente // lamenta // ~~un~~ dolore // acuto // ~~al~~ torace*
2. **Lemmatizzazione**
   1. ogni parola è ricondotta al lemma relativo
   2. esempio: *dolore*, *dolori* ⟶ *dolore*

Information flow del text preprocessing (3/4)

1. **Part of speech tagging**
   1. di ogni parola si indica la categoria grammaticale
   2. esempio: [dolore – N], [lamentare – V]
2. **Named entity recognition** (NER)
   1. la NER interessa nomi propri di persone, organizzazioni e luoghi; date e orari; valute, percentuali ecc.
   2. esempi: *Ippocrate*, *4 luglio p.v.*, *44%*
5. **Shallow/deep parsing**
   1. sulla base delle informazioni estratte e di un iniseme di pattern ricorrenti predefiniti, il sistema individua i sintagmi nominali e verbali o tutte le relazioni sintattiche
   2. esempi: [*il paziente scompensato* – SN], [*si sente male* – SV]

Information flow del text preprocessing (4/4)

5. **Building relations**
   1. le relazioni tra le entità individuate sono costruite attraverso pattern specifici
   2. esempio: [persona] [dichiara] [sintomo]
   3. ogni elemento del pattern è un etichetta per un insieme di forme linguistiche:
      [persona] = {paziente, soggetto, la signora X, il signor X…}
6. **Coreferenze**
   1. si individuano le relazioni di coreferenza
   2. esempio: **il paziente** è stato male e il medico **gli** ha detto che deve curar**si**
7. **Inferenze**
   1. il sistema può includere regole di inferenza per estrarre ulteriori informazioni

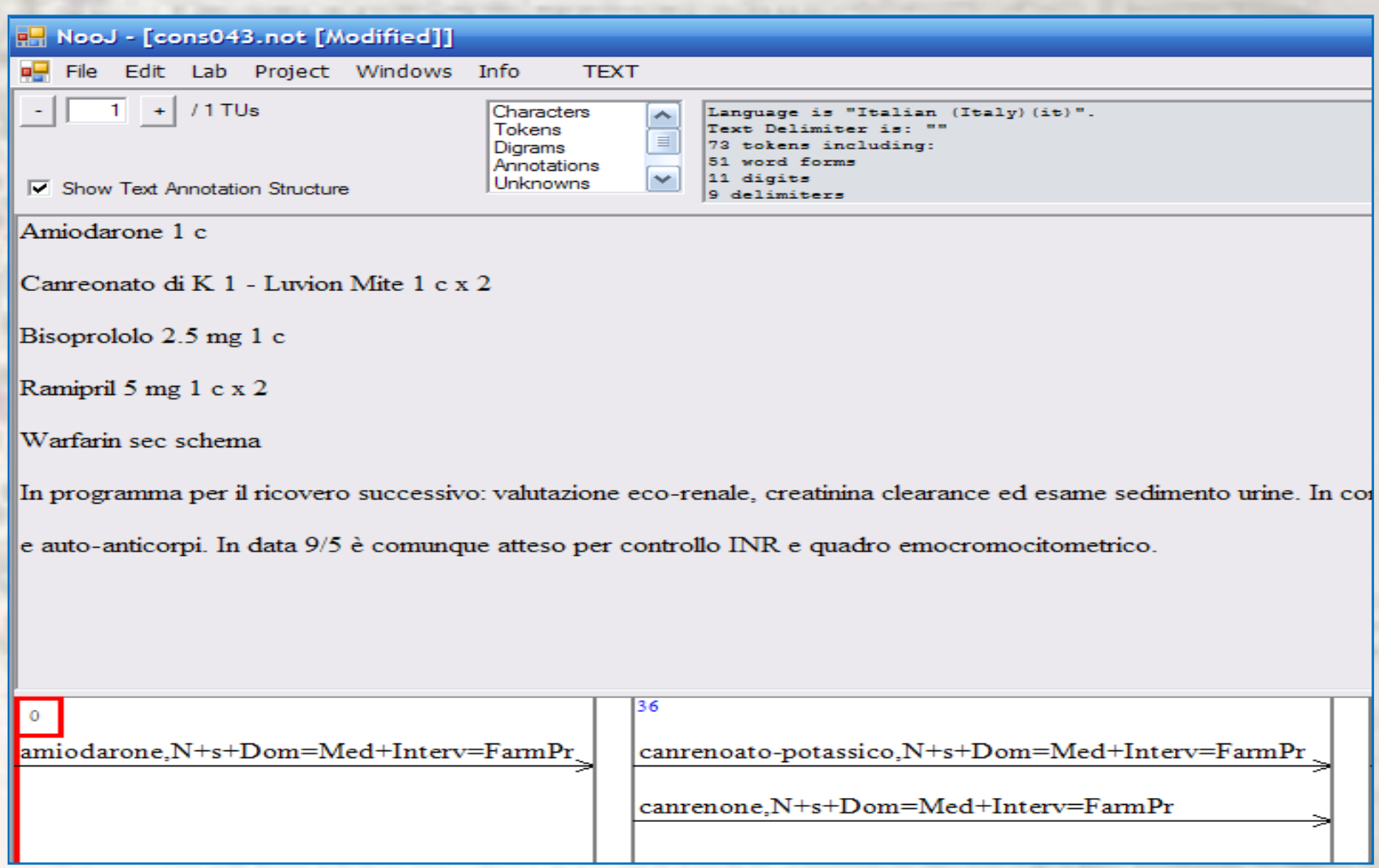
Il text preprocessing come NLP e IE

- Il text preprocessing si struttura come un processo di **natural language processing** e **information extraction**.
- Possiamo individuare **entità**, **concetti** e **correlazioni**:
  - nome = Ippocrate // categoria = persona
    nome = A. Osp. di Melegnano // categoria = organizzazione
  - lavora_presso[N-Y]
    lavora_presso[Ippocrate-A. Osp. di Melegnano].

Ontologie e dizionari

- Il text preprocessing richiede spesso il ricorso a **ontologie** e **dizionari**.
- Ontologia: sistema strutturato di concetti relativi al dominio in esame.
- Quando si lavora su un dominio semantico per uno scopo specifico, le basi di conoscenze sono decisive.

Esempio di applicazione di basi di conoscenze



Quali basi?

- **Standard** o costruite **ad hoc**
- **Generiche** (dizionari generici, WordNet) o **settoriali**
- Per il **settore biomedico**:
  - la versione più recente di UMLS (2007AA) integra 139 sorgenti terminologiche e ne trae un sistema concettuale codificato
  - Gene Ontology, UMLS Metathesaurus ecc.
- La scelta delle basi di conoscenza è sempre connessa all'applicazione del sistema e al dominio in esame.

Uso delle ontologie

- Le ontologie possono intervenire in momenti diversi

- **Uso attivo** delle ontologie: i concetti dell'ontologia informano la ricerca.

Problemi con le basi di conoscenze

- Costruire un dizionario o un'ontologia richiede **conoscenza esperta** e **lavoro umano**.
- Le basi di conoscenze non sono sempre **esportabili** verso altre applicazioni.
- L'**interoperabilità** non è garantita.

Il mining

«The core functionality of a text mining system resides in the analysis of concept co-occurrence patterns across documents in a collection». Feldman e Sanger (2007)

Che cosa si cerca

Pattern che si cercano nella fase di mining:

- **distribuzioni** di concetti e relazioni
- **frequent sets** e **near frequent sets** di testi per concetti e relazioni
- **associazioni** tra concetti e relazioni.

Text categorization

- La **text categorization** (TC) è la classificazione dei testi di una collezione in una serie di categorie predefinite.
- Usi di TC:
  - indicizzazione di testi sulla base di un vocabolario controllato
  - document sorting
  - filtraggio.
- Aspetti di TC:
  - **single-label** (ogni testo appartiene a una sola categoria) o **multilabel** (ogni testo può appartenere anche a più di una categoria)
  - per **decisioni binarie** o per **pesature**.

Clustering

- Il **clustering** è il raggruppamento dei testi di un corpus in categorie non predefinite, sulla base della **similarità** tra i testi stessi.
- Ipotesi di base: i testi rilevanti si assomigliano tra loro più che con quelli non rilevanti.

L’analisi di una collezione nel tempo

- **Analisi dei trend**: confronto tra sottoinsiemi cronologicamente definiti della collezione
- **Algoritmi incrementali**: aggiornamento progressivo dei risultati dell’analisi della collezione

L’interpretazione dei risultati

- L’interpretazione dei risultati avviene alla luce dell’**euristica** e degli **obiettivi** iniziali.
- Per agevolare l’interpretazione si possono raffinare i metodi di interrogazione del sistema e di visualizzazione dei risultati (GUIs).

Bibliografia (1/3)

- Ananiadou S., *Text Mining for Biomedicine: Techniques and tools*, The National Centre for Text Mining - The University of Manchester, scaricato da http://www.nactem.ac.uk/cs_teaching.php il 15 maggio 2009
- Bolasco S., Canzonetti S. e Caro M. F., *Text Mining: uno strumento strategico per imprese e istituzioni*, CISU, Roma, 2005
- Feldman R. e Sanger J., *The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data*, Cambridge University Press, New York, NY, 2007
- Dulli S., Polpettini P. e Trotta M. (a cura di), *Text mining: teoria e applicazioni*, FrancoAngeli, Milano, 2004
- Cineca, *Text Mining: aspetti applicativi in campo biomedico*, Cineca, 2001
- Cohen A. M., Hersh W. R., *A survey of current work in biomedical text mining*, «Briefings in Bioinformatics», 2005, vol. 6, n. 1, pp. 57-71

Bibliografia (2/3)

- de Bruijn B., Martin J., *Getting to the core of knowledge: mining biomedical literature*, «International Journal of Medical Informatics», 2002, vol. 67, n. 1-3, pp. 7-18
- Fiszman M., Chapman W., Aronsky D., Evans R. S., Haug P. J., *Automatic Detection of Acute Bacterial Pneumonia from Chest X-ray Reports*, «Journal of the American Medical Informatics Association», 2000, vol. 7, n. 6, pp. 593-604
- Forster A., Andrade J., van Walraven C., *Validation of a Discharge Summary Term Search Method to Detect Adverse Events,* «Journal of the American Medical Informatics Association», 2005, vol. 12, n. 2, pp. 200-206
- Hersh W., *Evaluation of biomedical text-mining systems: lessons learned from information retrieval*, «Briefings in bioinformatics», 2005, vol. 6, n. 4, pp. 344-356
- Hripcsak G., Austin J., Alderson P., Friedman C., *Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports,* «Radiology», 2002, vol. 224, n. 1, pp. 154-163

Bibliografia (3/3)

- Melton G., Hripcsak G., *Automated Detection of Adverse Events Using Natural Language Processing of Discharge Summaries*, «Journal of the American Medical Informatics Association», 2005, vol. 12, n. 4, pp. 448-457
- Murff H., Forster A., Peterson J., Fiskio J., Heiman H., Bates H., *Electronically Screening Discharge Summaries for Adverse Medical Events*, «Journal of the American Medical Informatics Association», 2003, vol. 10, n. 4, pp. 339-350
- Spasic I., Ananiadou S., McNaught J., Kumar A., *Text mining and ontologies in biomedicine: Making sense of raw text*, «Briefings in Bioinformatics», 2005, vol. 6, n. 3, pp. 239-251
- Swanson, D.R., *Complementary structures in disjoint science literatures*, in *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1991, ACM Press, Chicago, IL, pp. 280–289