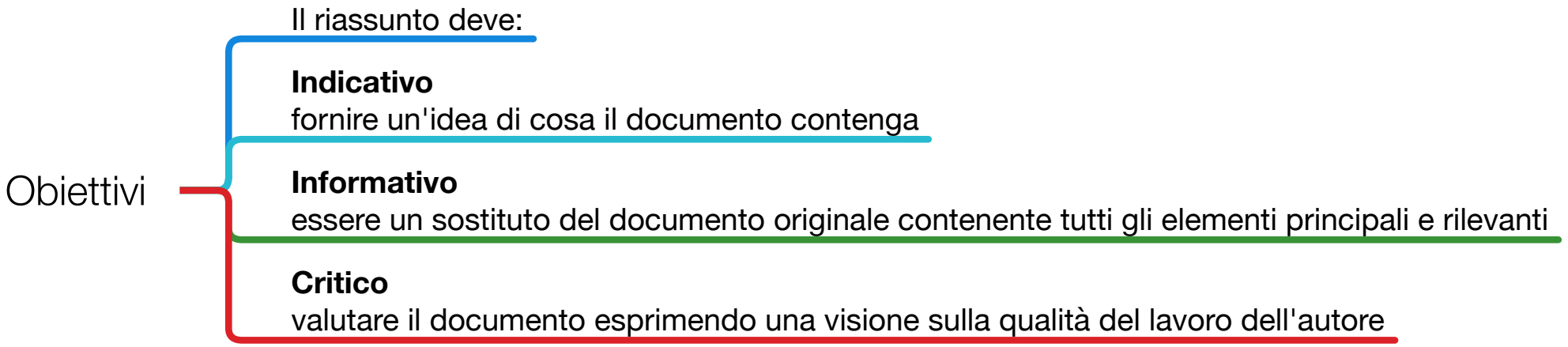
Riassunti automatici

Tutti questi elementi mostrati finora vengono utilizzati per la generazione automatica di riassunti.

L'obiettivo è di produrre un documento che risulti essere la sintesi di un documento di input oppure la sintesi e l'unione di due o più documenti.

Tale sintesi deve ovviamente contenere gli elementi rilevanti dei documenti iniziali.

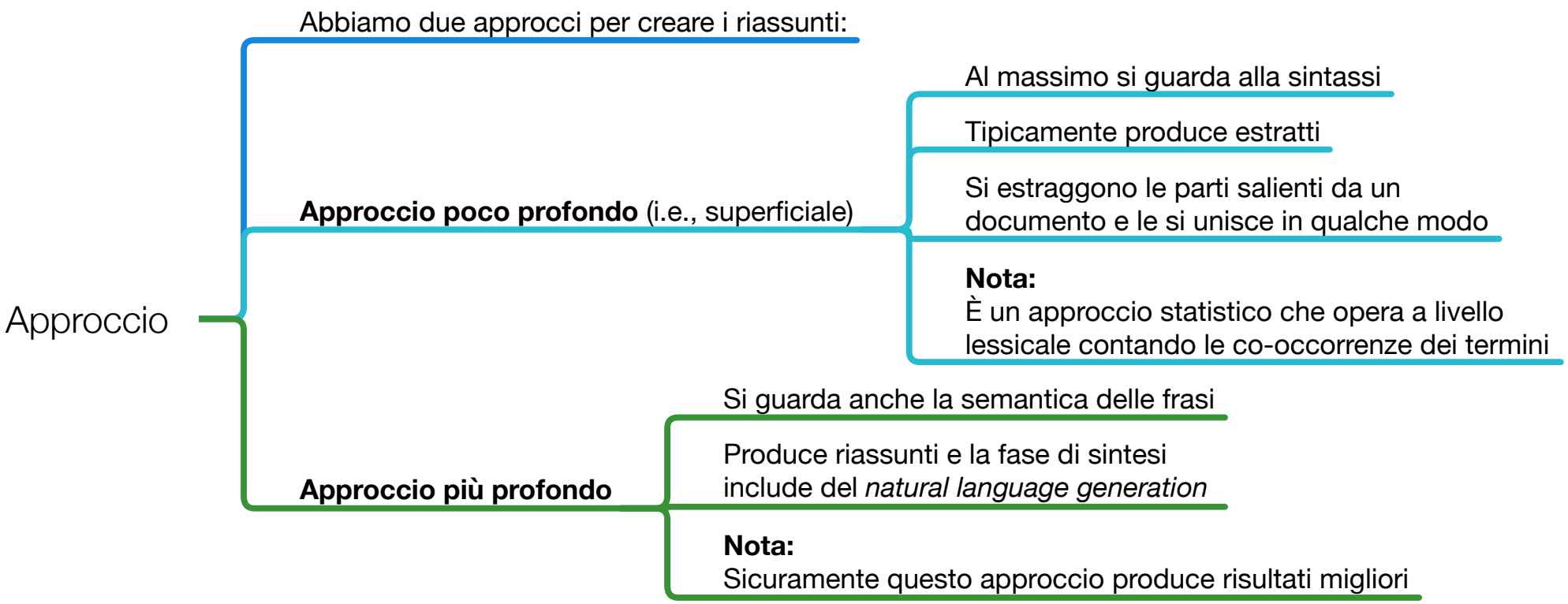

Obiettivi
Il riassunto deve:
Indicativo
fornire un'idea di cosa il documento contenga
Informativo
essere un sostituto del documento originale contenente tutti gli elementi principali e rilevanti
Critico
valutare il documento esprimendo una visione sulla qualità del lavoro dell'autore

Tipologie

I riassunti possono essere:

Estrattivi
Costruiti riutilizzando parti importanti del testo originario come frasi o paragrafi

Astrattivi
Che vanno a generare direttamente il riassunto in base alle informazioni raccolte senza riusarle in modo diretto (i.e., rielaborando il discorso)


Approccio
Abbiamo due approcci per creare i riassunti:
Approccio poco profondo (i.e., superficiale)
Al massimo si guarda alla sintassi
Tipicamente produce estratti
Si estraggono le parti salienti da un documento e le si unisce in qualche modo
Nota:
È un approccio statistico che opera a livello lessicale contando le co-occorrenze dei termini
Approccio più profondo
Si guarda anche la semantica delle frasi
Produce riassunti e la fase di sintesi include del *natural language generation*
Nota:
Sicuramente questo approccio produce risultati migliori

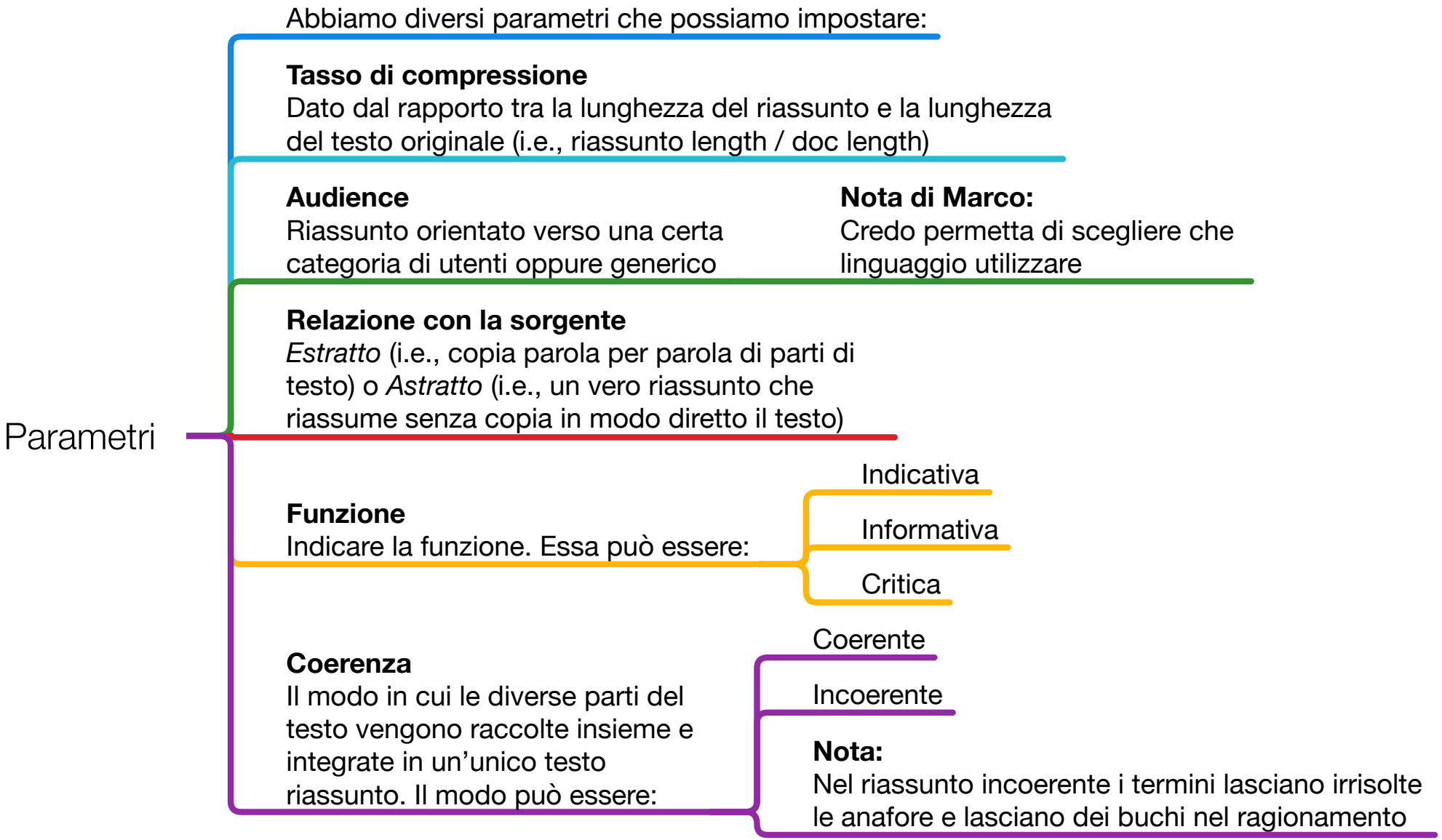
Riassunto
un doc
vs
tanti doc

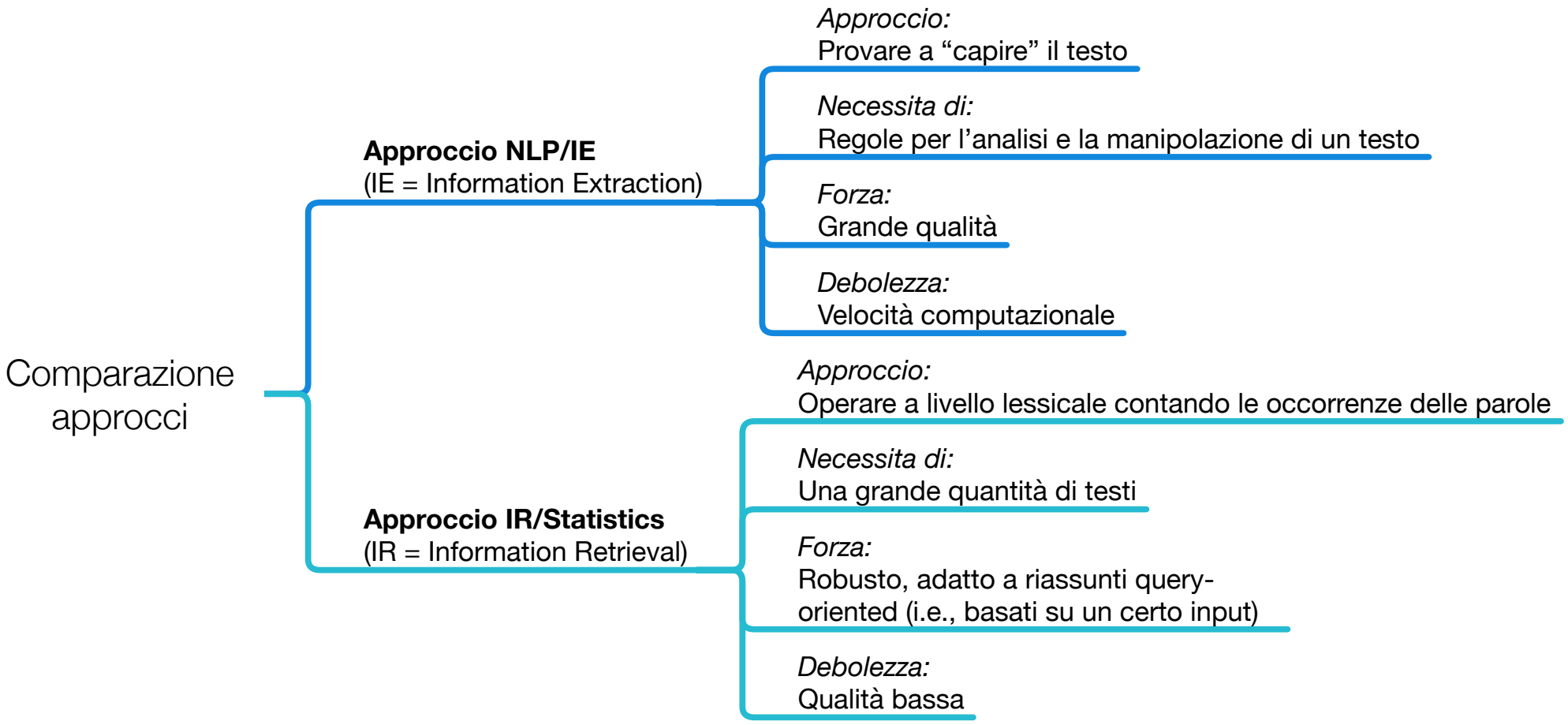
Riassunto singolo documento
Viene dato un singolo documento in input e viene prodotto il suo riassunto

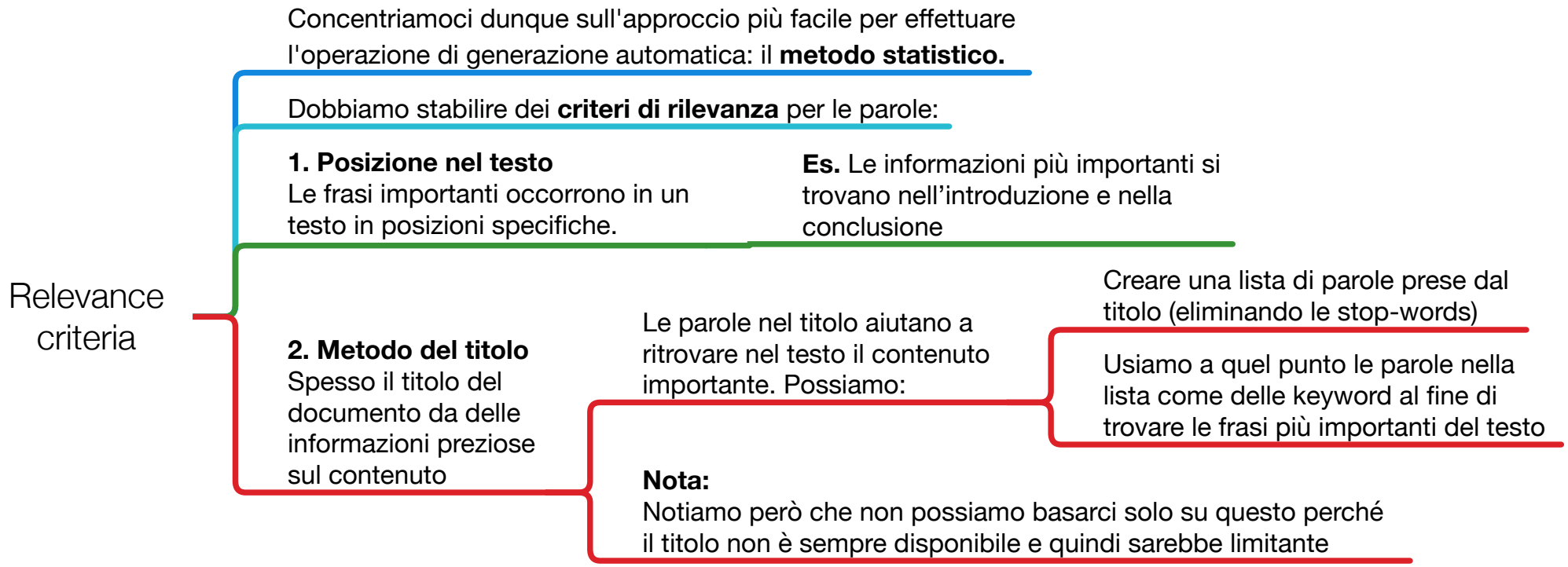
Tipicamente viene utilizzato in situazioni in cui si vuole produrre una **headline** o un'**outline**

Riassunto di tanti documenti insieme
In input abbiamo un gruppo di documenti e il nostro obiettivo è produrre un riassunto che è la condensazione del contenuto di tutto l'insieme di documenti

Tipicamente si usa per riassumere una serie di news sullo stesso evento oppure se abbiamo un qualsiasi contenuto web sullo stesso topic e lo vogliamo sintetizzare e condensare


Parametri
Abbiamo diversi parametri che possiamo impostare:
Tasso di compressione
Dato dal rapporto tra la lunghezza del riassunto e la lunghezza del testo originale (i.e., riassunto length / doc length)
Audience
Riassunto orientato verso una certa categoria di utenti oppure generico
Nota di Marco:
Credo permetta di scegliere che linguaggio utilizzare
Relazione con la sorgente
Estratto (i.e., copia parola per parola di parti di testo) o Astratto (i.e., un vero riassunto che riassume senza copia in modo diretto il testo)
Funzione
Indicare la funzione. Essa può essere:
Indicativa
Informativa
Critica
Coerenza
Il modo in cui le diverse parti del testo vengono raccolte insieme e integrate in un'unico testo riassunto. Il modo può essere:
Coerente
Incoerente
Nota:
Nel riassunto incoerente i termini lasciano irrisolte le anafore e lasciano dei buchi nel ragionamento


Comparazione approcci
Approccio NLP/IE
(IE = Information Extraction)
Approccio:
Provare a “capire” il testo
Necessita di:
Regole per l’analisi e la manipolazione di un testo
Forza:
Grande qualità
Debolezza:
Velocità computazionale
Approccio IR/Statistics
(IR = Information Retrieval)
Approccio:
Operare a livello lessicale contando le occorrenze delle parole
Necessita di:
Una grande quantità di testi
Forza:
Robusto, adatto a riassunti query-oriented (i.e., basati su un certo input)
Debolezza:
Qualità bassa


Relevance criteria
Concentriamoci dunque sull'approccio più facile per effettuare l'operazione di generazione automatica: il **metodo statistico.**
Dobbiamo stabilire dei **criteri di rilevanza** per le parole:
1. Posizione nel testo
Le frasi importanti occorrono in un testo in posizioni specifiche.
Es. Le informazioni più importanti si trovano nell'introduzione e nella conclusione
2. Metodo del titolo
Spesso il titolo del documento da delle informazioni preziose sul contenuto
Le parole nel titolo aiutano a ritrovare nel testo il contenuto importante. Possiamo:
Creare una lista di parole prese dal titolo (eliminando le stop-words)
Usiamo a quel punto le parole nella lista come delle keyword al fine di trovare le frasi più importanti del testo
Nota:
Notiamo però che non possiamo basarci solo su questo perché il titolo non è sempre disponibile e quindi sarebbe limitante

Relevance criteria

3. Optimum Position Policy (OPP)
La posizione delle frasi rilevanti è dipendente dal genere di documento.

Queste posizioni possono essere o conosciute a priori o determinate automaticamente attraverso degli algoritmi di learning

Esempio di algoritmo per determinare dove si trovano le frasi più importanti per un certo genere x:

- Estrarre la lista di keyword dal titolo
- Valutare in media dove si trovano le frasi che contengono le keyword (Optimal Position Policy)

4. Cue phrases method
Nei testi ci sono delle frasi che contengono parole che ci permettono di capire che stanno per essere dette cose importanti (i.e., bonus phrase) / inutili (i.e., stigma phrases)

Es. importanti:

- “The main aim of the present paper is to describe...”
- “The purpose of this article is to review...”
- “In this report, we outline...”
- “Our investigation has shown that...”

Nota:
Sono frasi che contengono comparativi, superlativi, espressioni conclusive, etc.

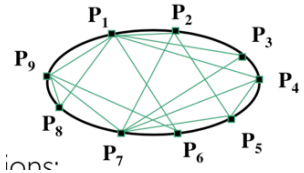
Es. inutili:

- Frasi che contengono “hardly”, “impossible”, etc

Queste frasi possono essere trovate automaticamente:

- Basta **aggiungere** ad una sentence (i.e., con sentence si intende una frase con verbo/predicato etc. una frase può anche non avere tutti questi elementi) dei punti se contiene *bonus frase* e toglierli se contiene *stigma frase*

Relevance criteria



5. Metodi basati sulla coesione
In generale le frasi importanti corrispondono alle entità maggiormente connesse all'interno di strutture semantiche.

Approcci
Si possono guardare:

- Le co-occorenze delle parole
- La co-reference (i.e., quando due o più espressioni in un testo si riferiscono alla stessa persona/cosa)
- Lexical similarity (WordNet)

Nota:
Sebbene questo metodo sia rozzo può essere utilizzato indipendentemente dalla lingua.

Coesione e co-occorrenza delle parole:

- Si può usare un classico metodo di IR dove si misura la similarità delle parole per determinare per ogni paragrafo P l'insieme S di paragrafi che sono in relazione con P
- Si determina lo score S per ogni paragrafo
- Si estraggono i paragrafi con gli score S più grandi

Descrizione sistemi auto summ

I sistemi di generazione automatica dei riassunti possono essere descritti in base a come risolvono i tre seguenti problemi:

1. Selezione del contenuto:
quali informazioni seleziona dai documenti

2. Ordinamento dell'informazione: come ordinano e strutturano l'informazione estratta

3. Realizzazione delle frasi:
che operazioni di pulizia vengono effettuate sulle frasi estratti (in modo tale che siano fluenti anche nel nuovo contesto/riassunto)

Algoritmo non supervisionato

In generale, l'algoritmo non-supervisionato più semplice che può produrre un riassunto:

- Prende uno o più criteri di rilevanza
- Seleziona le *sentence* che sono più importanti/informative/rilevanti
- Fissa dei valori minimi di importanza che queste *sentence* devono avere
- Va a prendere n paragrafi dal documento (quelli che superano i valori minimi di importanza)
- **Nota:**
  La rilevanza/importanza può essere misurata in diversi modi (e.g., word-frequency anche se una parola può essere molto probabile in Inglese ma non per quel particolare topic del documento)
- **Nota:**
  Decidere di quanto il documento riassunto deve essere più corto di quello originale (10% 20% 30%)

Es. algoritmo non supervisionato

1. Individua l'argomento del testo da riassumere. L'argomento può fare riferimento ad un insieme di vettori NASARI:

$$v_{t1} = \{term_1\_score, term_2\_score, \ldots, term_{10}\_score\}$$
$$v_{t2} = \{term_1\_score, term_2\_score, \ldots, term_{10}\_score\}$$
$$\ldots$$

2. Crea un contesto andando a raccogliere i vettori dei termini
3. Filtra i paragrafi tenendo quelli più rilevanti, ovvero quelli contenenti i termini più rilevanti

Nota:
Individuare i termini più rilevanti applicando almeno uno degli approcci sopra menzionati (i.e., relevance criteria) (e.g., title,cue,phrase,cohesion) e/o la nozione di *semantic similarity*