# 3 (to 1) steps

- Text summarisation systems are generally described by their solutions to the following three problems:
  - *Content Selection*: What information to select from the document(s) we are summarising. We usually make the simplifying assumption that the granularity of extraction is the sentence or clause. Content selection thus mainly consists of choosing which sentences or clauses to extract into the summary.
  - *Information Ordering*: How to order and structure the extracted units.
  - *Sentence Realisation*: What kind of clean up to perform on the extracted units so they are fluent in their new context.

# unsupervised algorithm

- The simplest unsupervised algorithm is to select sentences that have more salient or informative words.
  - Sentences that contain more informative words tend to be more extract-worthy.
- *Saliency* is usually defined by computing the topic signature, a set of salient or signature terms, each of whose saliency scores is greater than some threshold $\theta$.
  - Saliency could be measured in terms of simple word frequency, but frequency has the problem that a word might have a high probability in English in general but not be particularly topical to a particular document.
- Lexical specificity can thus be adopted in order to individuate the most salient terms, and to score the sentences where they appear.

# a simple *extractive* algorithm

- reduce the document size of e.g., 10%, 20%, 30%

1. individuate the topic of the text being summarised; the topic can be referred to as a (set of) NASARI vector(s):

   $v_{t1}$ = {*term*$_1$*_score, term*$_2$*_score, …, term*$_{10}$*_score* }
   $v_{t2}$ = {*term*$_1$*_score, term*$_2$*_score, …, term*$_{10}$*_score* }
   …

2. create the context, by collecting the vectors of terms herein (this step can be repeated, by dumping the contribution of the associated terms at each round);

3. retain paragraphs whose sentences contain the most salient terms, based on the Weighted Overlap, *WO*($v_1$,$v_2$)
   - rerank paragraphs weight by applying at least one of the mentioned approaches (*title*, *cue*, *phrase*, *cohesion*).

# NASARI (lexical) subset

- two distribution files are provided for NASARI, that require different resources allocation.
  - *dd-nasari.txt*. a subset of NASARI (obtained by truncating vectors at 10 features). 3,587,754 vectors, ~600MB;

    https://goo.gl/85BubW
  - *dd-small-nasari-15.txt*. a subset of NASARI. same filtering as above, with 15 features + intersection with 60K lemmas in the Corpus of Contemporary American English: 13,084 vectors, 2MB storage (many entities removed here…).
- the second one has been extracted for starting our experimentation; the second one is intended to explore the resource in a richer (though reduced) flavour.

# documents for summarisation

- text documents are provided for summarisation purposes:
    - *Andy-Warhol.txt*
    - *Ebola-virus-disease.txt*
    - *Life-indoors.txt*
    - *Napoleon-wiki.txt*
- do experiment with different compression rates: 10%, 20% and 30%.