Text as data

DSTA

Big picture

- [verbal] language is the main vehicle of human comunication
- until short ago, the main encoding of data/knowledge

. . .

- languages represent what we value and [seems to] determine what we **can** think.
- universal grammars are studied.

- complex grammars: a sign of affectation?
- terms are ambiguos **by design**

. . .

- the polar opposite of what data analytics needs!

From the British Medical Journal:

Organ specific immune-related adverse events are uncommon with anti-PD-1 drugs but the risk is increased compared with control treatments. General adverse events related to immune activation are largely similar. Adverse events consistent with musculoskeletal problems are inconsistently reported but adverse events may be common.

From Men’s Health:

The stats don’t lie... here’s how to get the looks of her dreams

Love is in the eye of the beholder.

But good looks are down to science... sort of.

Ah, the chest. The part of the body most men would like to grow. Luckily, you have come to the right place. We really do know a thing or two about building muscle. Take your pick from the workouts below to stretch your chest.

From Text to numerical methods

The occurrence matrix

Often text (document) analysis begins with word occurrence analysis: we record word usage irrespective of the position in text.

. . .

Let a document be a bag of words. Let a corpus be a collection of documents.

. . .

The occurence matrix A:

$a_{ij} = k$

means that word i appears k times in document j.

Word i is represented by the i-th row of A (also A_i^T)

. . .

With row normalisation, word usage becomes a prob. distribution to which Entropy analysis can be applied.

With column norm., we analyse a document by its entropy.