2. Term-Document incidence matrices

an **incidence matrix** is a matrix that shows the relationship between two classes of objects.

Unstructured data in 1620

- which plays of Shakespeare contain the words **Brutus** AND **Caesar** but NOT **Calpurnia**?
- One could grep all of Shakespeare's plays for **Brutus** and **Caesar**, then strip out lines containing **Calpurnia**?
- Why is the not the answer?
  - Slow : corpora의 크기가 커질수록 느려짐
  - NOT **Calpurnia** is non-trivial : **Calpurnia**을 포함하지 않는 문서를 처리하는 것이 쉽지 않음
  - Other operations (e.g., find the word **Romans** near **countrymen**) not feasible : **countrymen** 단어 주위에 **Romans** 단어가 있는 문서를 찾을 수가 없음
  - Ranked retrieval (best documents to return) : 문서를 ranking한 것이 아님

Term-Document incidence matrices

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

`Brutus AND caesar BUT NOT Calpurnia`

- 1, if play contains word
- 0, otherwise

Term-Document matrix를 이용하면 boolean query를 query로 입력할 때 search engine이 result를 반환하기 쉽다.

Incidence vectors

- So we have a 0/1 vector for each term
- To answer query : take the vectors for **Brutus**, **Caesar** and **Calpurnia** (complemented) → bitwise AND.
  - 110100 (**Brutus**) AND
  - 110111 (**Caesar**) AND
  - 101111 (not **Calpurnia**) =
  - 100100
  - Antony and Cleopatro, Hamlet이 query에 만족하는 play

- → Term-Document matrix를 이용해 간단한 information retrieval 가능

Bigger collections

- consider N = 1 million documents, each with about 1000 words.
- Avg 6 bytes/word including spaces/punctuation
    - 6GB of data in the docuemnts
- Say there are M = 500K **distinct** terms among these

Cant's build the matrix

- 500K × 1M matrix has five hundred billion 0's and 1's → 크기가 너무 커져 Term-Document matrix를 만들기 어렵
- But it has no more than one billion 1's (한 문서는 약 1000개의 단어들로 구성되기에 (1,000,000 × 1,000)) → 그마저도 불용어들이 많음 → **sparse matrix**
- What's a better representation ?
    - We only record the 1 positions

출처 : https://en.wikipedia.org/wiki/Incidence_matrix

https://www.youtube.com/watch?v=e81nC0LO0A8&list=PLaZQkZp6WhWwoDuD6pQCmgVyDbUWl_ZUi&index=2