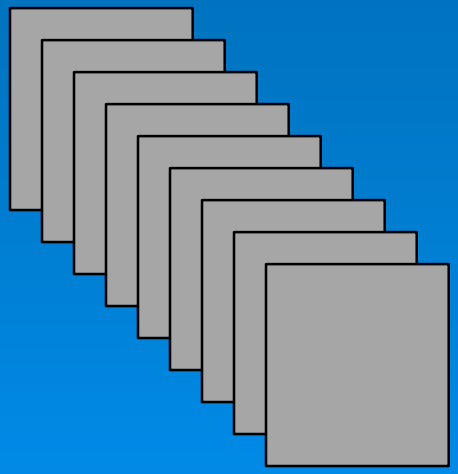
DOCUMENT RELATIONSHIP ANALYSIS

A presentation by
W H Inmon

Document relationship analysis

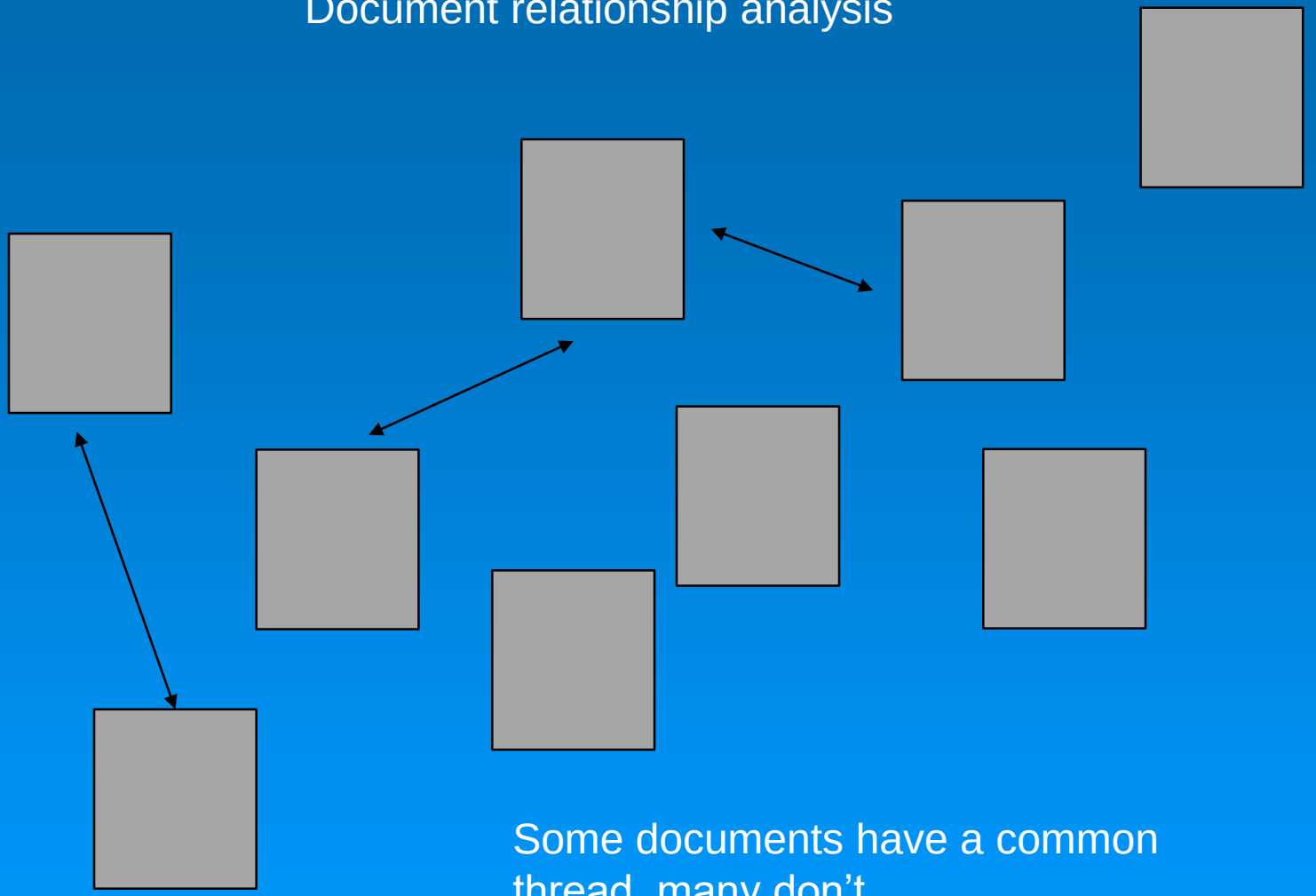


Suppose you have a lot of documents

Suppose you need to find a common thread connecting certain documents

FOREST RIM TECHNOLOGY INC

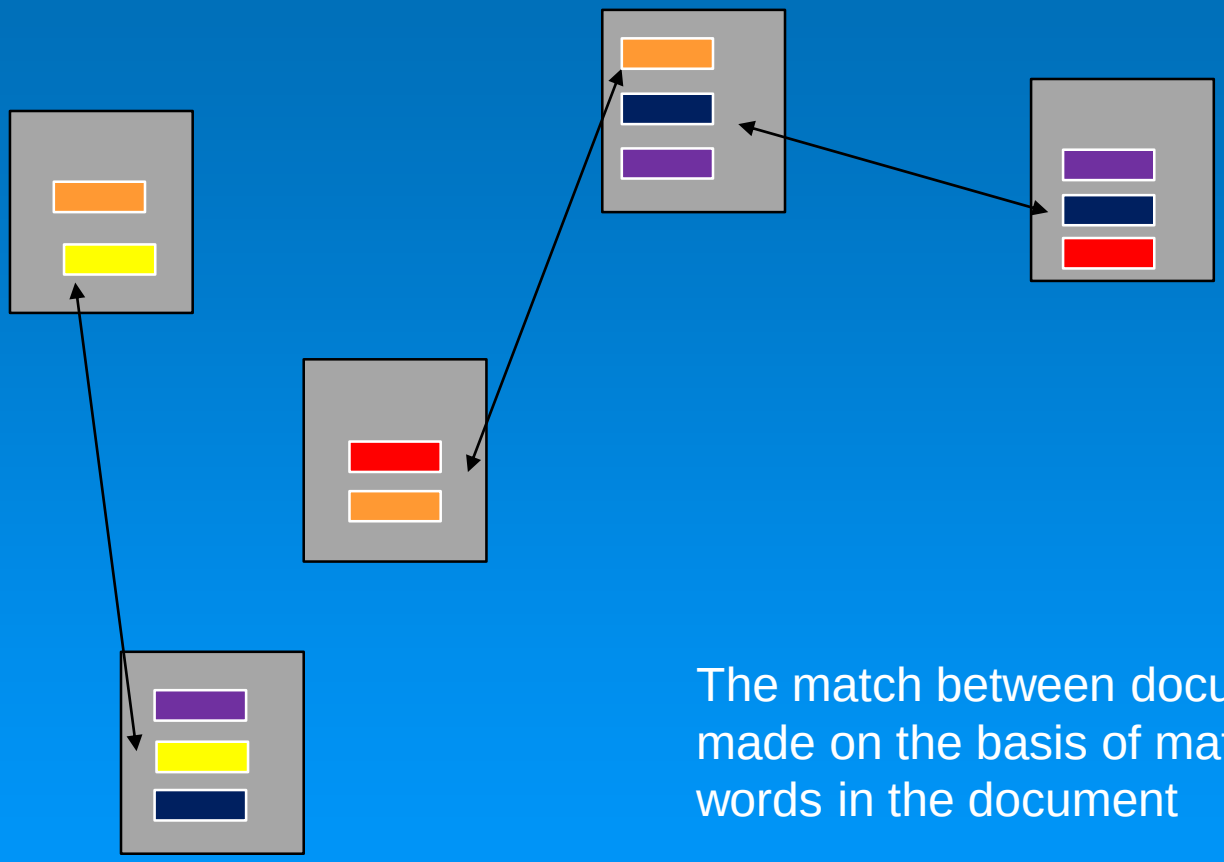
Document relationship analysis



Some documents have a common thread, many don’t

FOREST RIM TECHNOLOGY INC

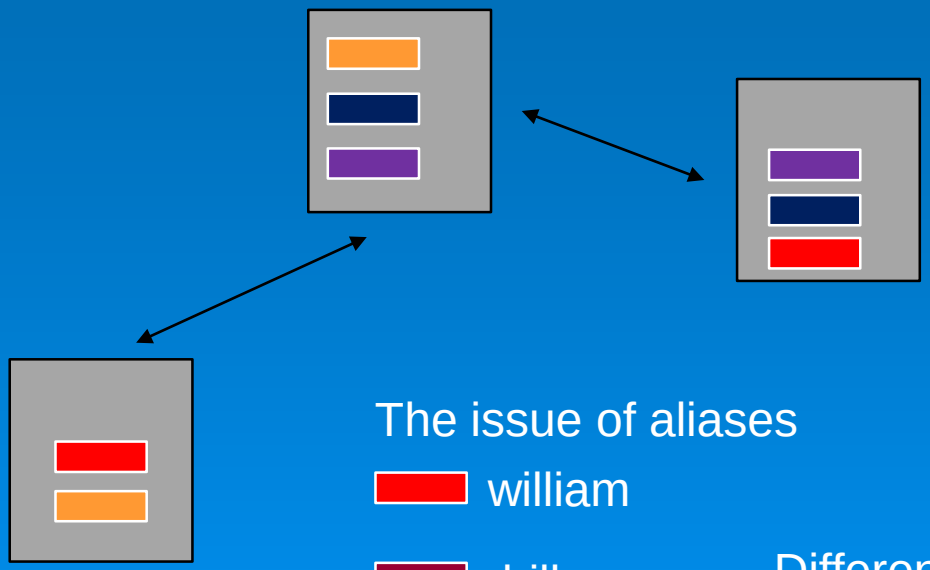
Document relationship analysis



The match between documents is made on the basis of matching words in the document

FOREST RIM TECHNOLOGY INC

Document relationship analysis



The issue of aliases

william

bill

willie

billy

Different names for the same person

A complicating issue that must be taken into account

FOREST RIM TECHNOLOGY INC

Document relationship analysis

lasix

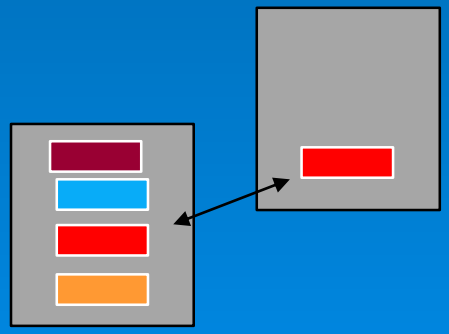
furosemide

Di-aqua-2

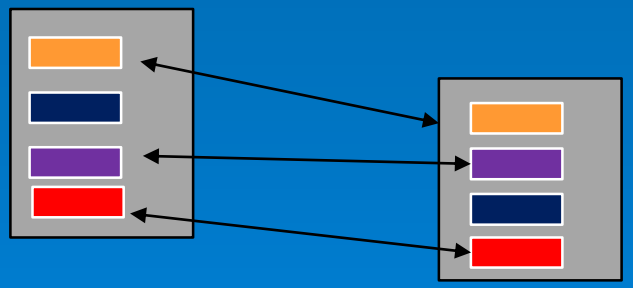
All three are the same thing.
But they are hardly spelled the same

FOREST RIM TECHNOLOGY INC

Document relationship analysis



A weak match

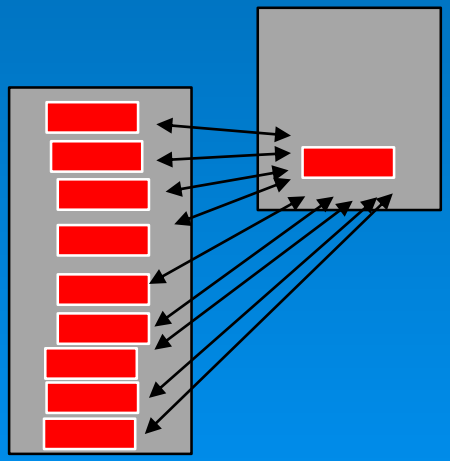


A strong match

Another issue –
the strength of the match

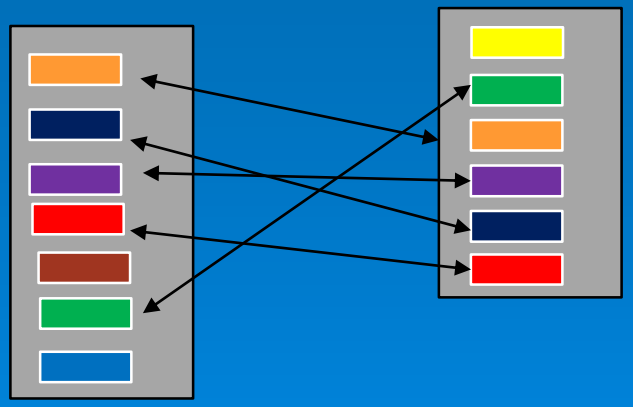
FOREST RIM TECHNOLOGY INC

Document relationship analysis



Total matching words – 10
Ratio – 9:1

A weak match



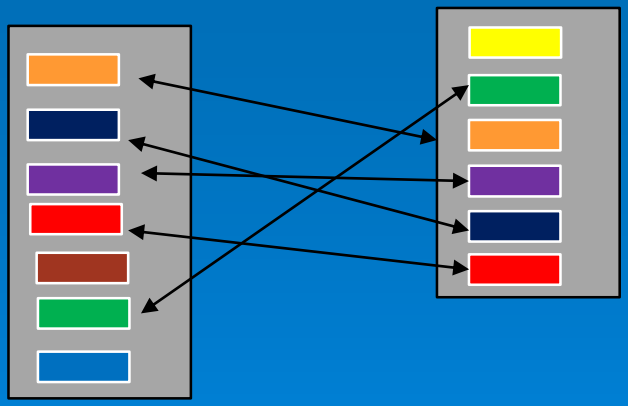
Total matching words – 10
Ratio - 5:5

A strong match

The issue of the ratio of words has a big effect on measuring the strength of the relationship

FOREST RIM TECHNOLOGY INC

Document relationship analysis



A – matching words
In document 1

B – matching words
In document 2

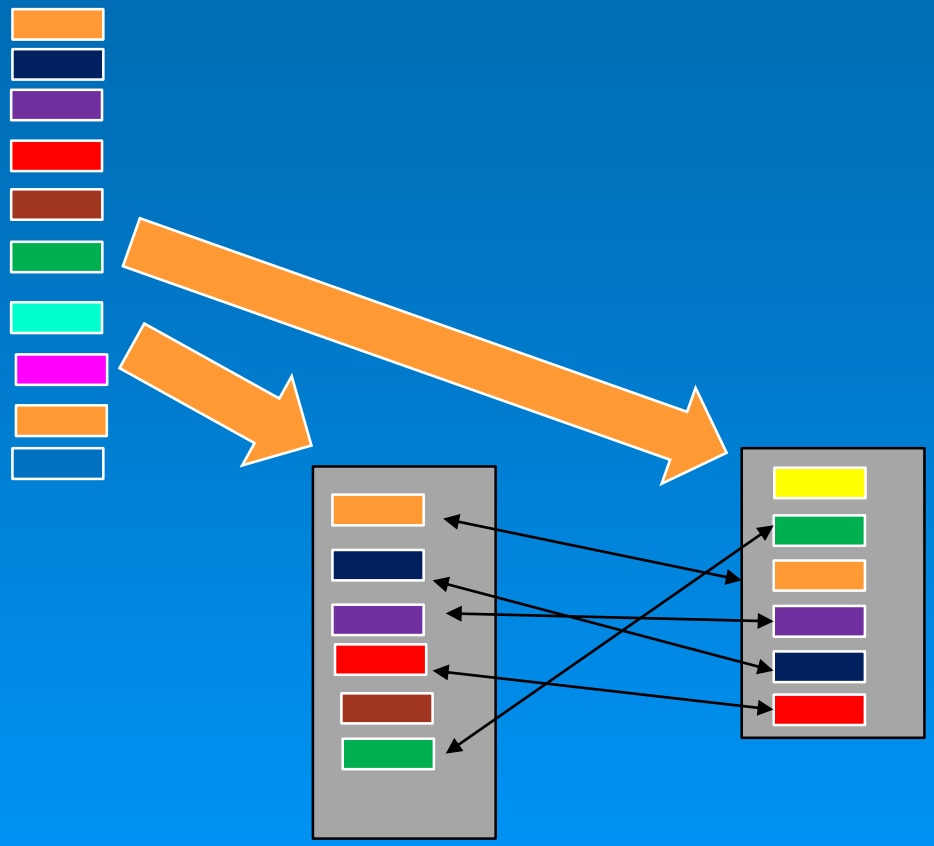
Measuring the strength of the match –

$$\sum_{N=1,2,3\ldots} (a + b) \times ((\min, a,b) /(\max(a,b))$$

For all matches between a and b

FOREST RIM TECHNOLOGY INC

Document relationship analysis

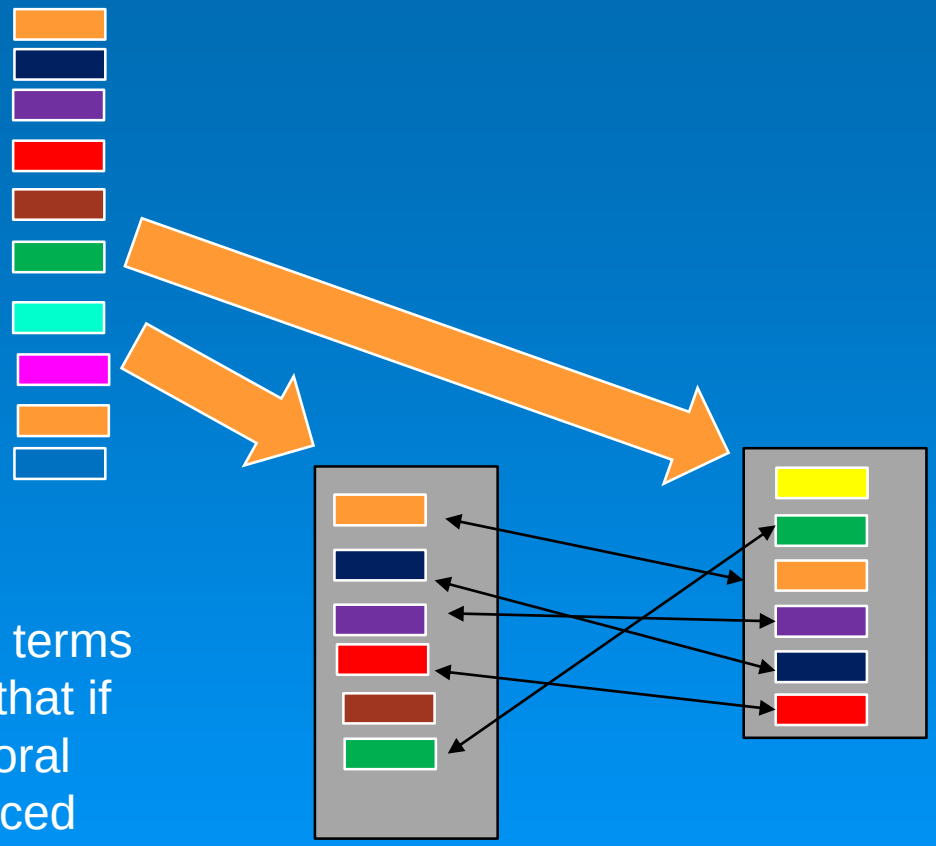


The behavioral issue

FOREST RIM TECHNOLOGY INC

Document relationship analysis

Need to be flexible in selecting words that express behavior

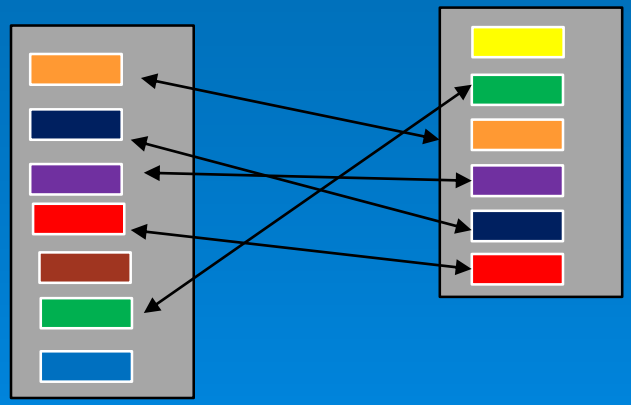


Important behavioral terms may be obscure, so that if needed, new behavioral terms can be introduced

The behavioral issue

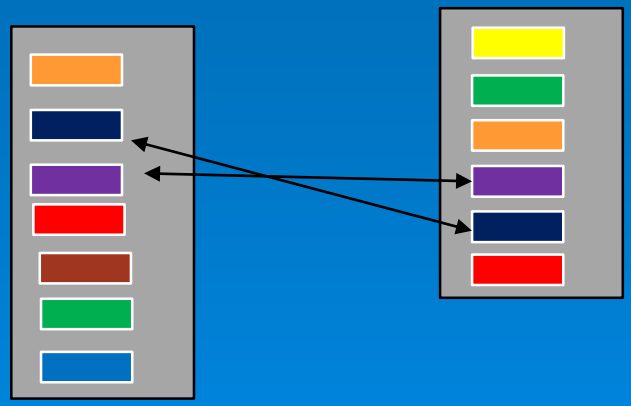
FOREST RIM TECHNOLOGY INC

Document relationship analysis



All word analysis

“word by word”



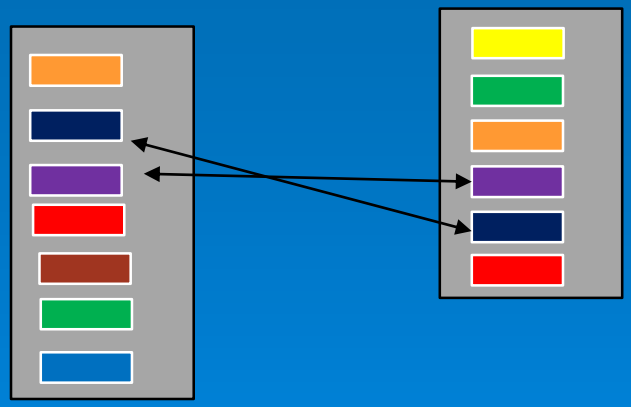
Selected word analysis

“composite analysis”

There are two very different types of analysis
that can be done to analyze the strength of the match

FOREST RIM TECHNOLOGY INC

Document relationship analysis



"composite analysis"

A composite is made up of more than one term

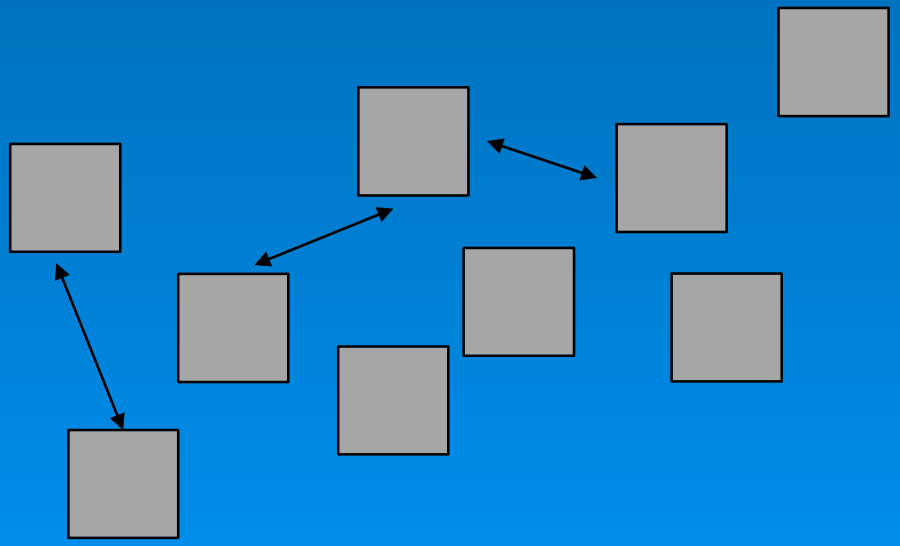
Terms –
- white
- bronco

Or
- Heisman
- USC
- running back

A composite analysis is done based ONLY on the words found in the composite

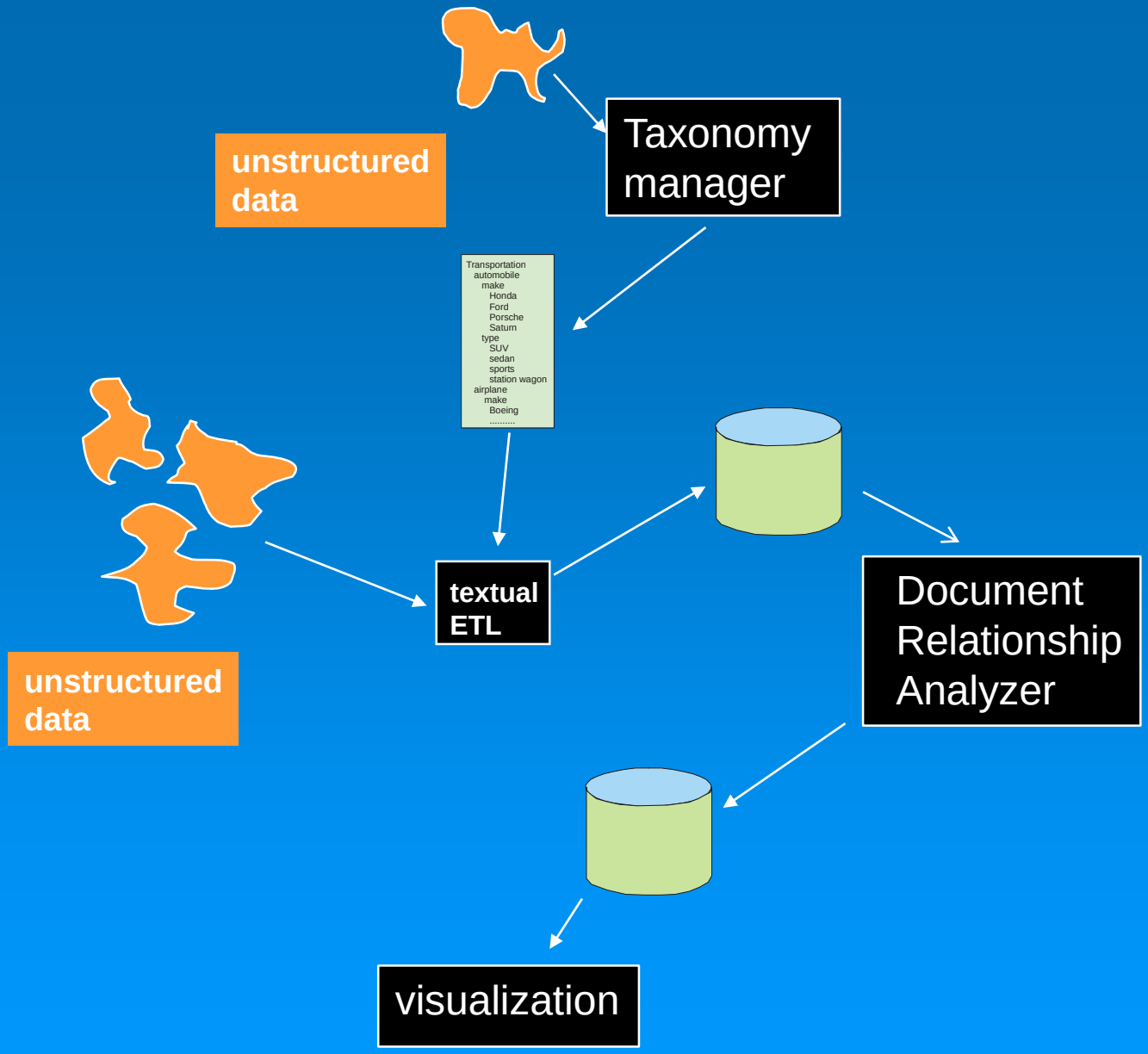
FOREST RIM TECHNOLOGY INC

Document relationship analysis

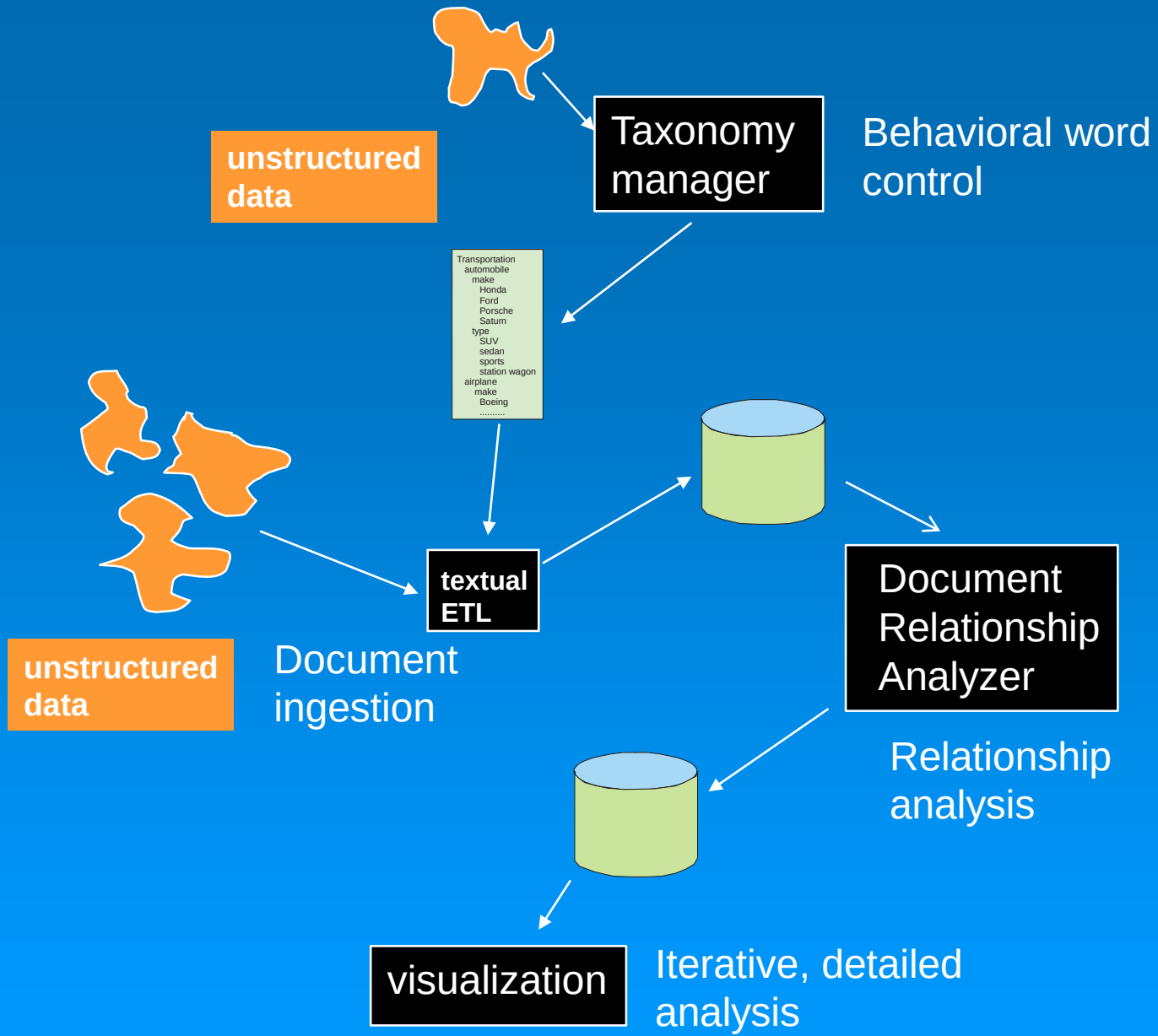


There are many facets to matching documents

FOREST RIM TECHNOLOGY INC



FOREST RIM TECHNOLOGY INC



FOREST RIM TECHNOLOGY INC

OJ Simpson

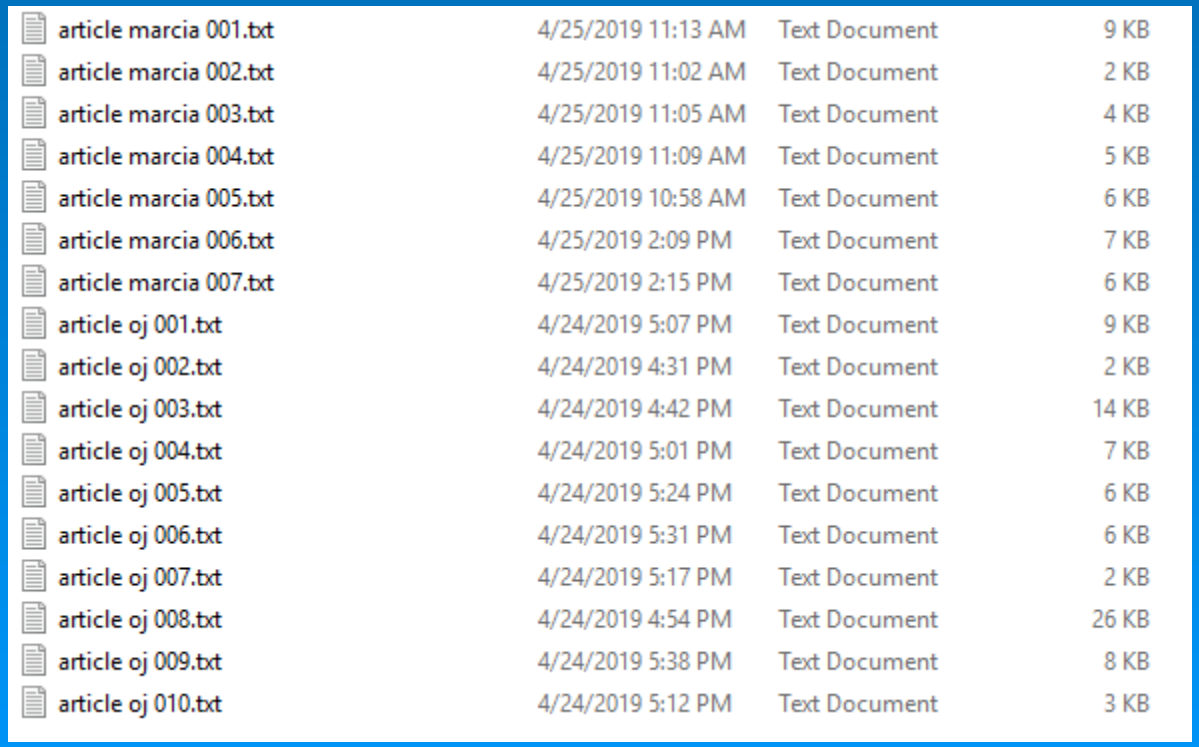
Marcia Clark

A simple example
The OJ Simpson trial

Widely available articles in the news

FOREST RIM TECHNOLOGY INC

The articles that were stripped from the news



article marcia 001.txt	4/25/2019 11:13 AM	Text Document	9 KB
article marcia 002.txt	4/25/2019 11:02 AM	Text Document	2 KB
article marcia 003.txt	4/25/2019 11:05 AM	Text Document	4 KB
article marcia 004.txt	4/25/2019 11:09 AM	Text Document	5 KB
article marcia 005.txt	4/25/2019 10:58 AM	Text Document	6 KB
article marcia 006.txt	4/25/2019 2:09 PM	Text Document	7 KB
article marcia 007.txt	4/25/2019 2:15 PM	Text Document	6 KB
article oj 001.txt	4/24/2019 5:07 PM	Text Document	9 KB
article oj 002.txt	4/24/2019 4:31 PM	Text Document	2 KB
article oj 003.txt	4/24/2019 4:42 PM	Text Document	14 KB
article oj 004.txt	4/24/2019 5:01 PM	Text Document	7 KB
article oj 005.txt	4/24/2019 5:24 PM	Text Document	6 KB
article oj 006.txt	4/24/2019 5:31 PM	Text Document	6 KB
article oj 007.txt	4/24/2019 5:17 PM	Text Document	2 KB
article oj 008.txt	4/24/2019 4:54 PM	Text Document	26 KB
article oj 009.txt	4/24/2019 5:38 PM	Text Document	8 KB
article oj 010.txt	4/24/2019 5:12 PM	Text Document	3 KB

FOREST RIM TECHNOLOGY INC

At the end of a sensational trial, former football star O.J. Simpson is acquitted of the brutal 1994 double murder of his estranged wife, Nicole Brown Simpson, and her friend, Ronald Goldman. In the epic 252-day trial, Simpson's "dream team" of lawyers employed creative and controversial methods to convince jurors that Simpson's guilt had not been proved "beyond a reasonable doubt," thus surmounting what the prosecution called a "mountain of evidence" implicating him as the murderer.

Orenthal James Simpson–a Heisman Trophy winner, star running back with the Buffalo Bills, and popular television personality–married Nicole Brown in 1985. He reportedly regularly abused his wife and in 1989 pleaded no contest to a charge of spousal battery. In 1992, she left him and filed for divorce. On the night of June 12, 1994, Nicole Brown Simpson and Ronald Goldman were stabbed and slashed to death in the front yard of Mrs. Simpson's condominium in Brentwood, Los Angeles. By June 17, police had gathered enough evidence to charge O.J. Simpson with the murders.

Simpson had no alibi for the time frame of the murders. Some 40 minutes after the murders were committed, a limousine driver sent to take Simpson to the airport saw a man in dark clothing hurrying up the drive of his Rockingham estate. A few minutes later, Simpson spoke to the driver though the gate phone and let him in. During the previous 25 minutes, the driver had repeatedly called the house and received no answer.

A single leather glove found outside Simpson's home matched a glove found at the crime scene. In preliminary DNA tests, blood found on the glove was shown to have come from Simpson and the two victims. After his arrest, further DNA tests would confirm this finding. Simpson had a wound on his hand, and his blood was a DNA match to drops found at the Brentwood crime scene. Nicole Brown Simpson's blood was discovered on a pair of socks found at the Rockingham estate. Simpson had recently purchased a "Stiletto" knife of the type the coroner believed was used by the killer. Shoe prints in the blood at Brentwood matched Simpson's shoe size and later were shown to match a type of shoe he had owned. Neither the knife nor shoes were found by police.

On June 17, a warrant was put out for Simpson's arrest, but he refused to surrender. Just before 7 p.m., police located him in a white Ford Bronco being driven by his friend, former teammate Al Cowlings. Cowlings refused to pull over and told police over his cellular phone that Simpson was suicidal and had a gun to his head. Police agreed not to stop the vehicle by force, and a low-speed chase ensued. Los Angeles news helicopters learned of the event unfolding on their freeways, and live television coverage began. As millions watched, the Bronco was escorted across Los Angeles by a phalanx of police cars. Just before 8 p.m., the dramatic journey ended when Cowlings pulled into the Rockingham estate. After an hour of tense negotiation, Simpson emerged from the vehicle and surrendered. In the vehicle was found a travel bag containing, among other things, Simpson's passport, a disguise kit consisting of a fake moustache and beard, and a revolver. Three days later, Simpson appeared before a judge and pleaded not guilty.

Simpson's subsequent criminal trial was a sensational media event of unprecedented proportions. It was the longest trial ever held in California, and courtroom television cameras captured the carnival-like atmosphere of the proceedings. The prosecution's mountain of evidence was systemically called into doubt by Simpson's team of expensive attorneys, who made the dramatic case that their client was framed by unscrupulous and racist police officers. Citing the questionable character of detective Mark Fuhrman and alleged blunders in the police investigation, defense lawyers painted Simpson as yet another African American victim of the white judicial system. The jurors' reasonable doubt grew when the defense spent weeks attacking the damning DNA evidence, arguing in overly technical terms that delays and other anomalies in the gathering of evidence called the findings into question. Critics of the trial accused Judge Lance Ito of losing control of his courtroom.

In polls, a majority of African Americans believed Simpson to be innocent of the crime, while white America was confident of his guilt.

An article about OJ Simpson

FOREST RIM TECHNOLOGY INC

The OJ Simpson taxonomy

news agency	daily news	store	null	null
news agency	espn	store	null	null
news agency	getty images	store	null	null
news agency	vanity fair	store	null	null
news agency	watching newsletter	store	null	null
nfl	n f l	store	null	null
nfl	n. f. l.	store	null	null
nfl	nfl	store	null	null
nickname	the juice	store	null	null
oj simpson	o j simpson	store	null	null
oj simpson	o. j. simpson	store	null	null
oj simpson	o.j. simpson	store	null	null
oj simpson	oj simpson	store	null	null
oj simpson	orenthal	store	null	null
oj simpson	orenthal james simpson	store	null	null
parole	parole	store	null	null
passport	passport	store	null	null
person	al "a c " cowlings	store	null	null
person	al cowlings	store	null	null
person	alan dershowitz	store	null	null
person	alfred beardsley	store	null	null

FOREST RIM TECHNOLOGY INC

Finding OJ

news agency	daily news	store	null	null
news agency	espn	store	null	null
news agency	getty images	store	null	null
news agency	vanity fair	store	null	null
news agency	watching newsletter	store	null	null
nfl	n f l	store	null	null
nfl	n. f. l.	store	null	null
nfl	nfl	store	null	null
nickname	the juice	store	null	null
oj simpson	o j simpson	store	null	null
oj simpson	o. j. simpson	store	null	null
oj simpson	o.j. simpson	store	null	null
oj simpson	oj simpson	store	null	null
oj simpson	orenthal	store	null	null
oj simpson	orenthal james simpson	store	null	null
parole	parole	store	null	null
passport	passport	store	null	null
person	al "a c " cowlings	store	null	null
person	al cowlings	store	null	null
person	alan dershowitz	store	null	null
person	alfred beardsley	store	null	null

FOREST RIM TECHNOLOGY INC

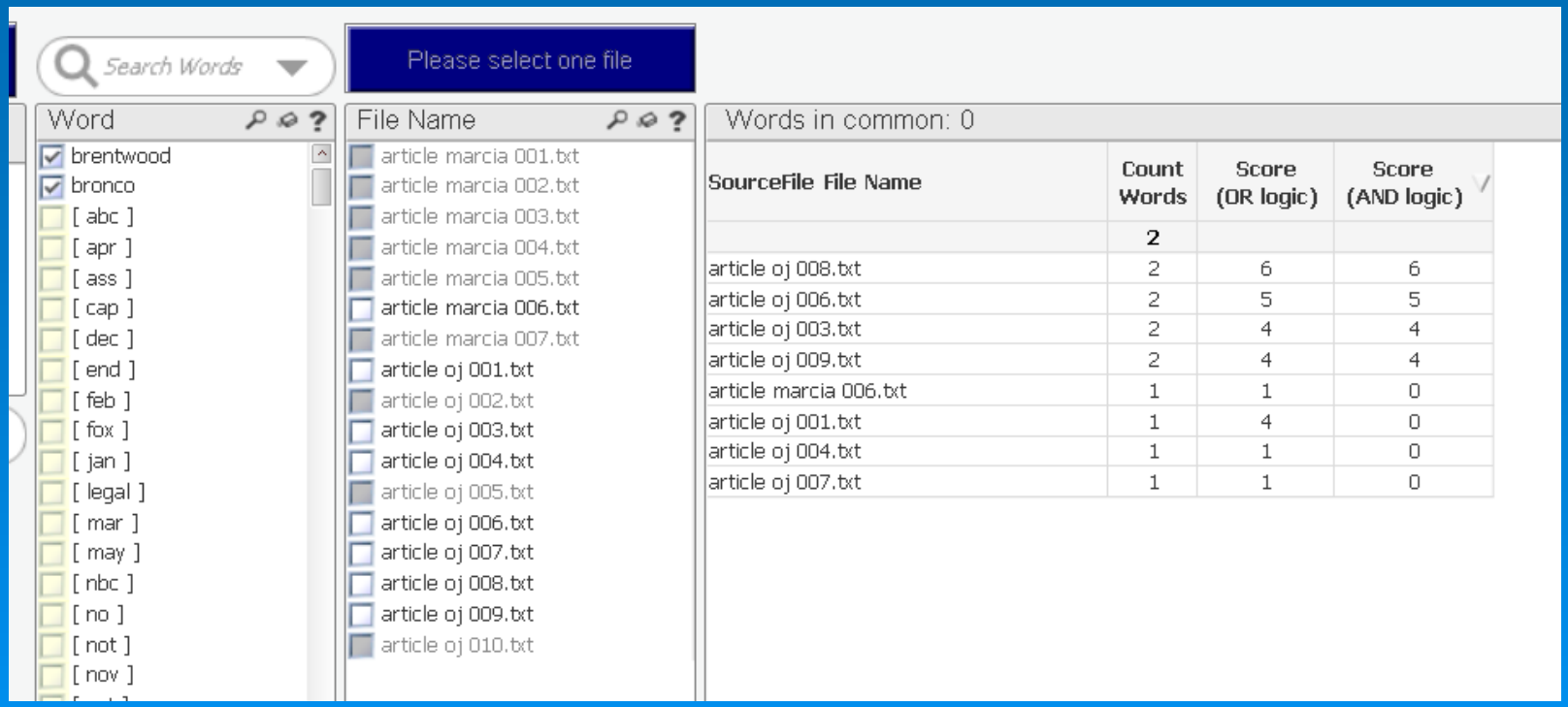
Document 1	Document 2	Combined Matched words		Composite Name
C:\proof of co - OJ Simpson\article marcia 003.txt	C:\proof of conce J Simpson\article marcia 004...	16	16	composite z
C:\proof of co - OJ Simpson\article marcia 003.txt	C:\proof of conce J Simpson\article marcia 005...	16	16	composite z
C:\proof of co - OJ Simpson\article marcia 003.txt	C:\proof of conc J Simpson\article marcia 006...	16	16	composite tz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article marcia 007...	16	16	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 001.txt	16	16	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 002.txt	16	16	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 003.txt	16	16	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 004.txt	16	13	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 005.txt	16	13	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 006.txt	16	13	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 007.txt	16	13	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 008.txt	16	13	composite - hertz
C:\proof of concept - OJ Simpson\article marcia 003.txt	C:\proof of concept - OJ Simpson\article oj 009.txt	16	13	composite - hertz

The composite data base

FOREST RIM TECHNOLOGY INC

Word by word analysis

Article 8 and article 9 have the strongest relationship



Analytical spreadsheet for matching documents

FOREST RIM TECHNOLOGY INC

So who needs a document relationship analysis?

Police departments looking to manage internal documents
Police departments looking to compare documents with other police departments
Pharmaceuticals looking to analyze test results
Medical research
Insurance claims looking for fraud
Many others……

People who have a lot of documents
People who have unstructured textual based documents
People who need to relate a collection of documents to a "foreign" collection of documents

FOREST RIM TECHNOLOGY INC