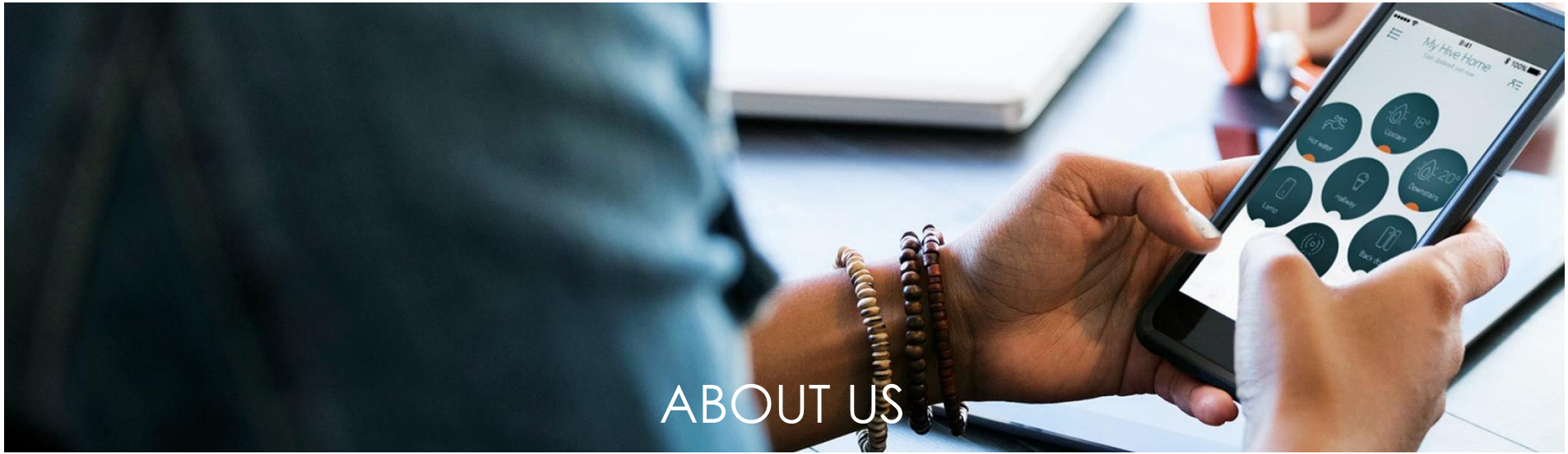
centrica

KölnR

Köln R User Group (15th December 2017)

Text Mining for Preventative Maintenance

Timothy Wong
Senior Data Scientist, Centrica plc



ABOUT US

We supply energy and services to over 27 million customer accounts

Supported by around 12,000 engineers and technicians

Our areas of focus are Energy Supply & Services, Connected Home, Distributed Energy & Power, Energy Marketing & Trading

Bord Gáis Energy

British Gas

centrica

Direct Energy

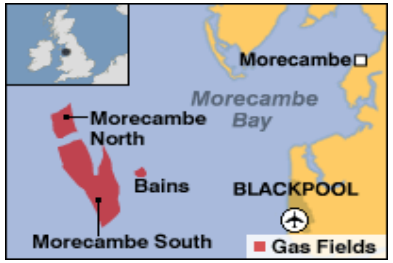
DYNO

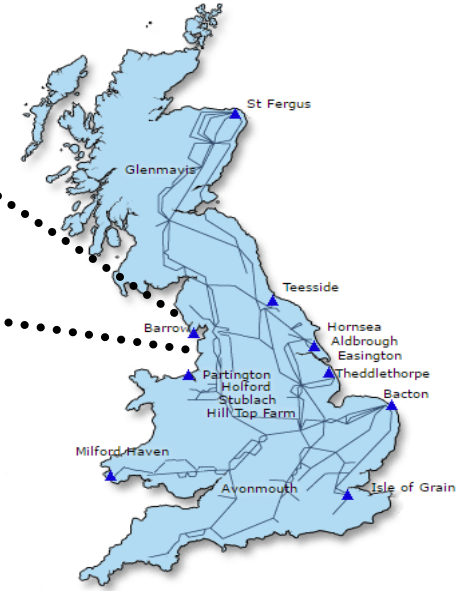
HIVE

Local Heroes

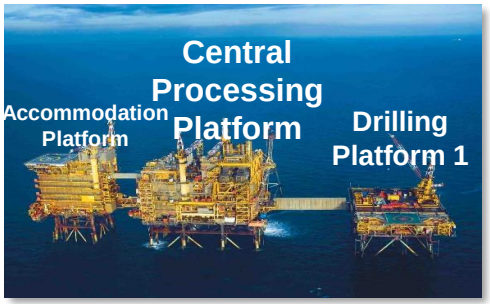
Energy Exploration & Production

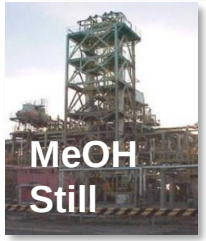
centrica





Approx. 6% of GB's gas supply





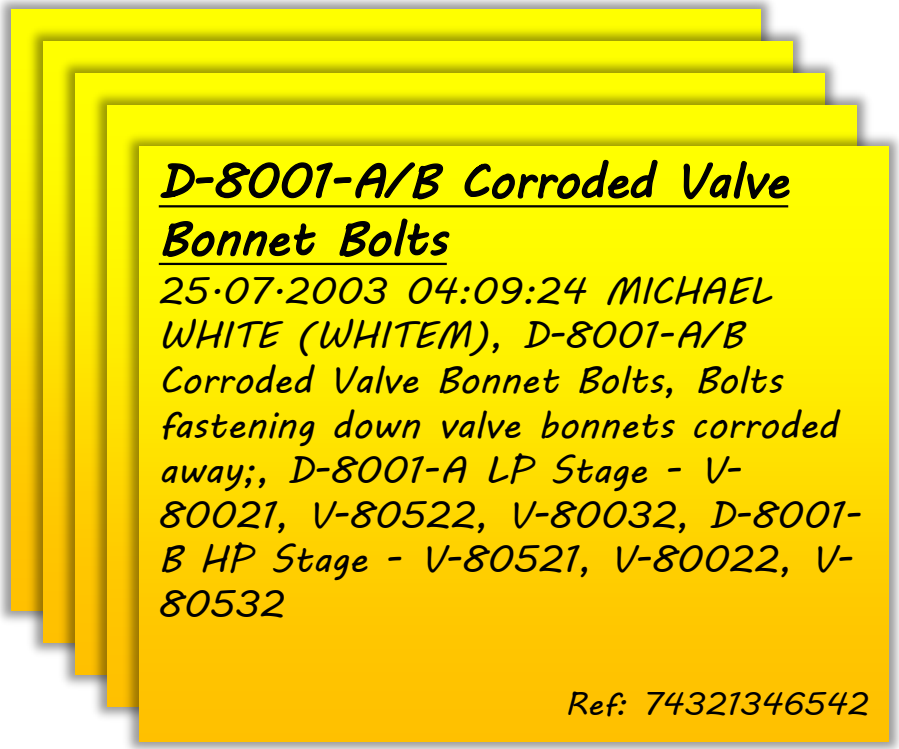
The Morecambe Terminals

centrica

- East Irish Sea
  - Maintenance events/incidents are recorded in the repair log
  - Identify recurring vulnerabilities
  - Analytics approach - Text mining algorithms (Natural Language Processing)

Source Data

centrica

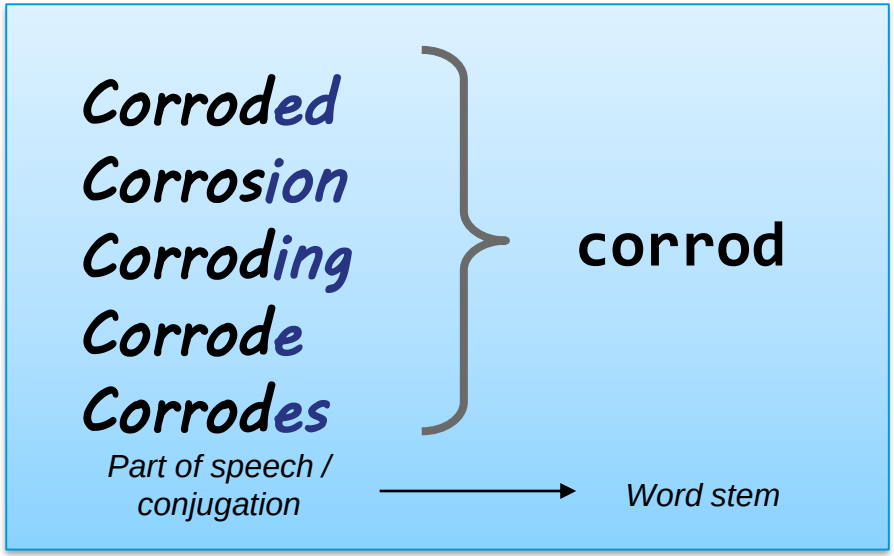


- Repair log system
  - Unstructured text data
  - Lots of technical abbreviations
    - Component IDs
    - Locations IDs
    - Names, datetime… etc
  - Incomplete syntax
  - Sometimes having typos

Term Extraction (1)

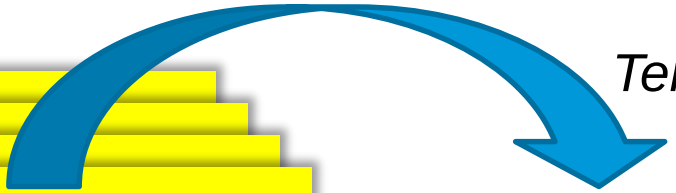
centrica

1. Covert to lower case
2. Remove non-alphabets
3. Remove common words
4. Word stemming



Term Extraction (2)

centrica



D-8001-A/B Corroded Valve Bonnet Bolts

~~25·07·2003 04:09:24 MICHAEL WHITE (WHITEM),~~ D~~-8001-~~A~~/~~B Corrod~~ed~~ Valve Bonnet Bolt~~s,~~ Bolt~~s~~ fasten~~ing~~ ~~down~~ valve bonnet~~s~~ corrod~~ed~~ away~~;,~~ D~~-8001-~~A LP Stage = V~~-80021,~~ V~~-80522,~~ V~~-80032,~~ D~~-8001-~~B HP Stage = V~~-80521,~~ V~~-80022,~~ V~~-80532~~

Ref: 74321346542

D-8001-A/B Corroded Valve Bonnet Bolts

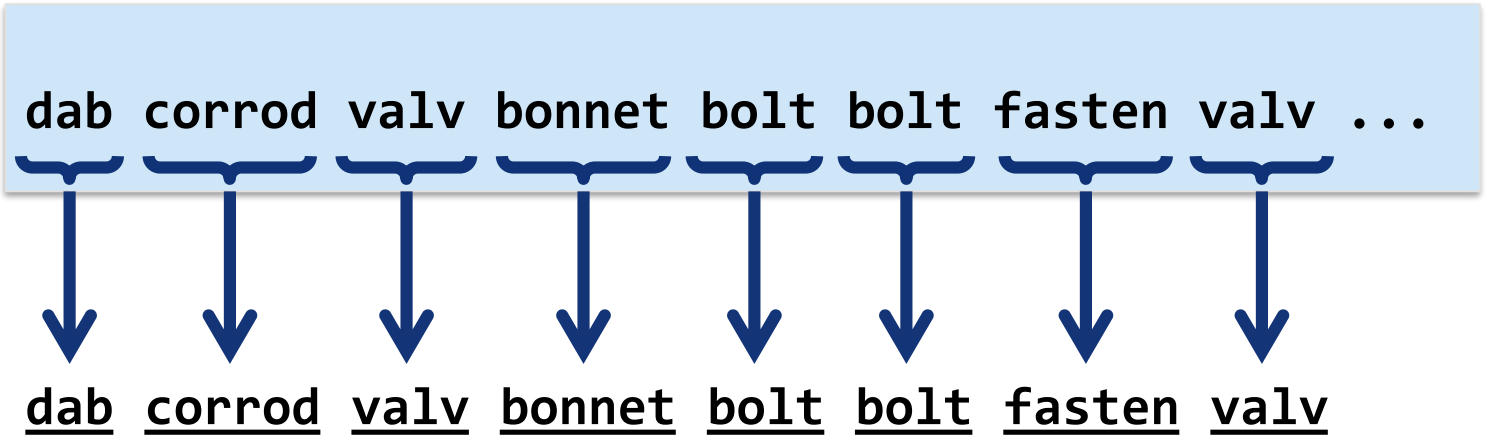
dab corrod valv bonnet bolt bolt fasten valv bonnet corrod away da lp stage v v v db hp stage v v v

Ref: 74321346542

Term Extraction (3): n-gram model

centrica

- Unigram ($n = 1$)

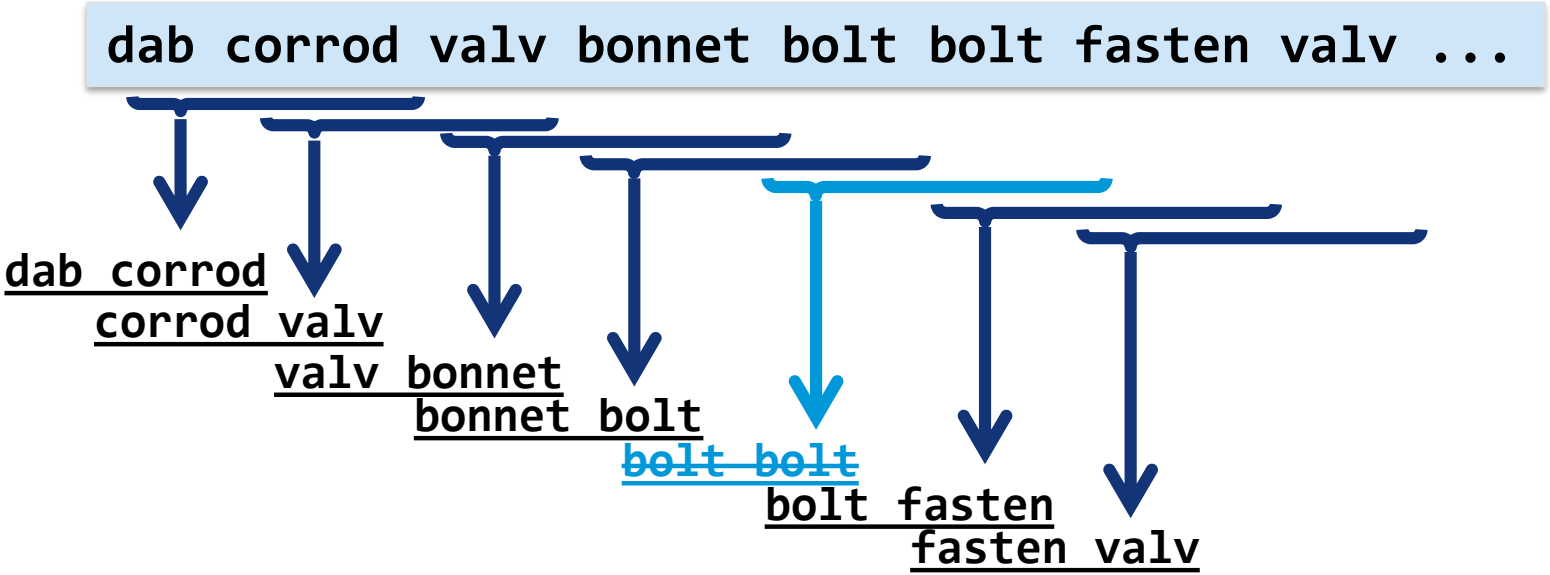


Term Extraction (4): n-gram model

centrica

- Bigram ($n = 2$)

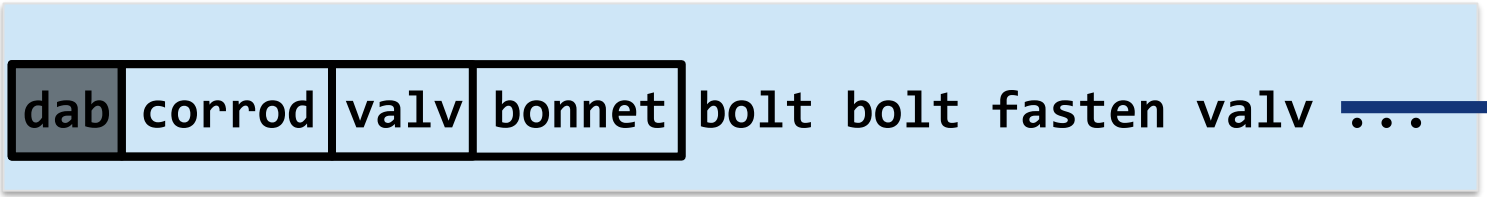
  Terms with repetitive words are removed.



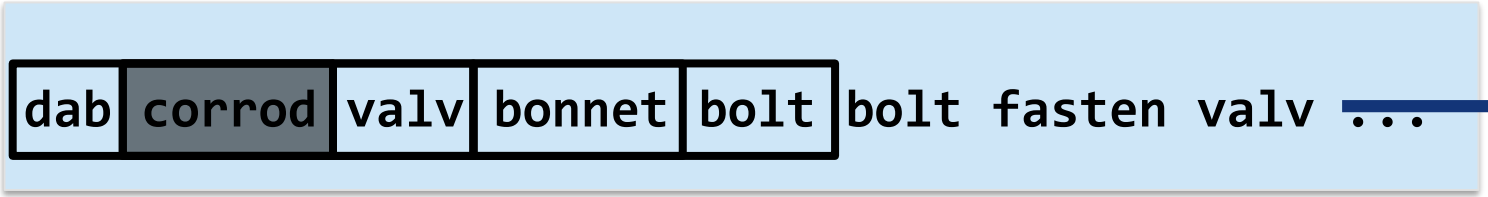
Term Extraction (5): skip-gram model

centrica

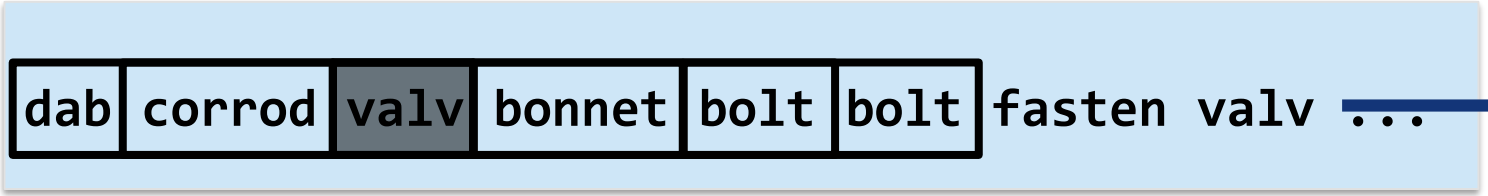
- Uses a rolling window and takes pair of words



dab corrod
dab valv
dab bonnet



corrod dab
corrod valv
corrod bonnet
corrod bolt



valv dab
valv corrod
valv bonnet
valv bolt
valv bolt

Information Retrieval: tf-idf scheme (1)

centrica

- **Term Frequency (tf)**

  Reflects occurrence of term t in a given document d

$$tf_{t,d} = \frac{Number\ of\ occurance\ of\ term\ t\ in\ document\ d}{Total\ number\ of\ terms\ in\ document\ d}$$

- **Inverse Document Frequency (idf)**

  Reflects occurrence of term in entire corpus D

$$idf_{t,D} = \log\left(\frac{Number\ of\ documents\ in\ corpus\ D}{Number\ of\ documents\ having\ term\ t}\right)$$

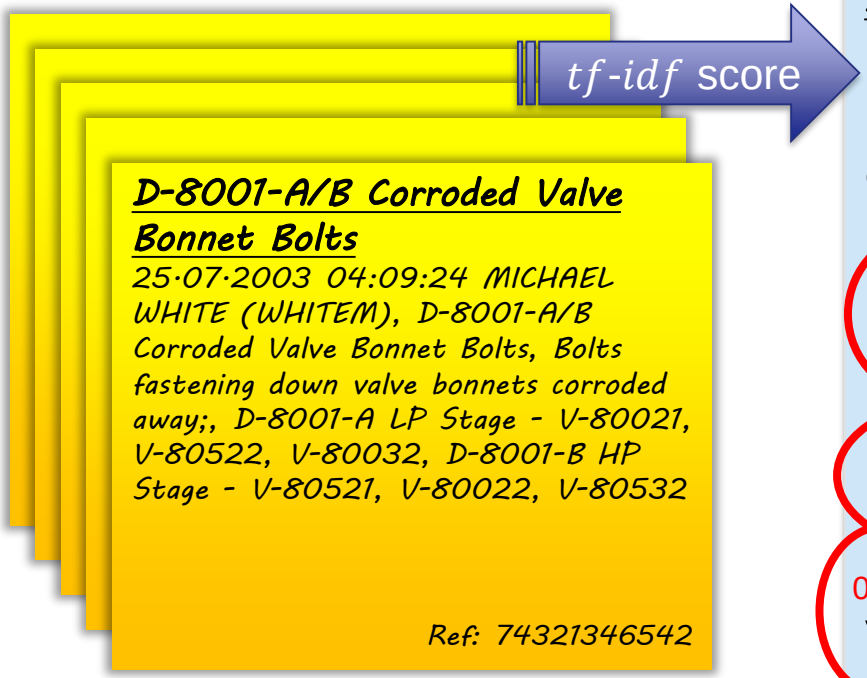
- **Weighted Term Importance ($tf{-}idf$)**

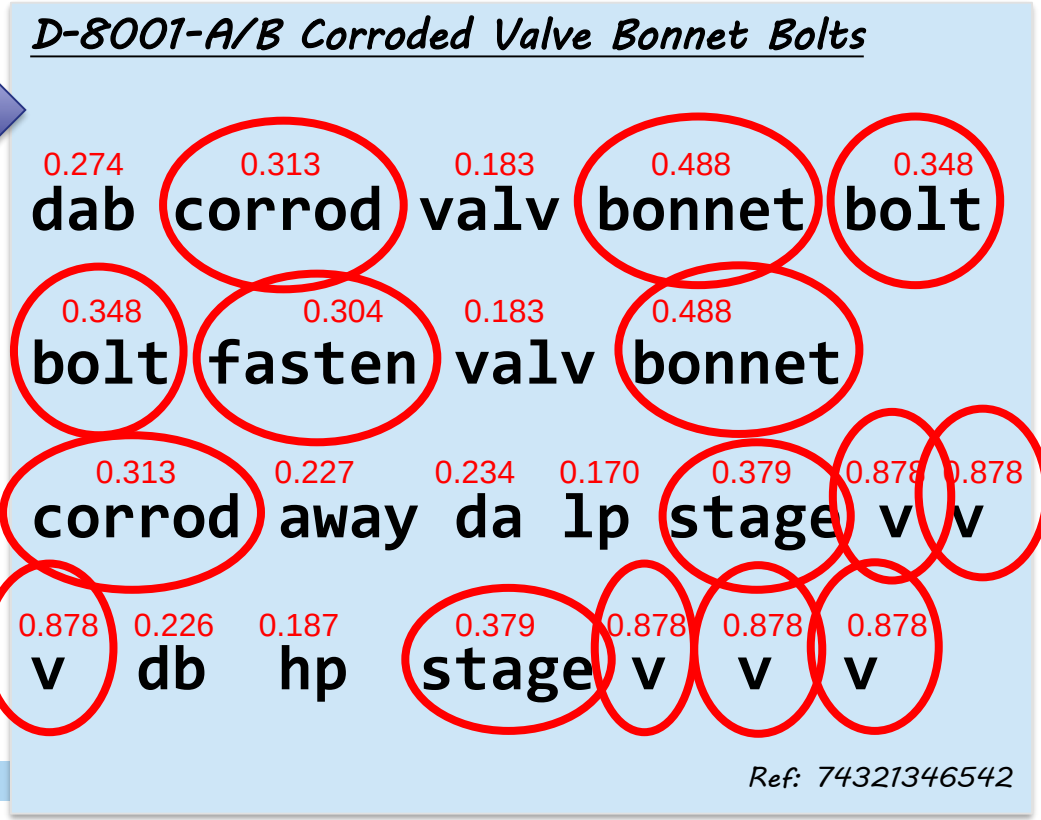
$$tfidf = tf_{t,d} \times idf_{t,D}$$

Information Retrieval: tf-idf scheme (2)

centrica

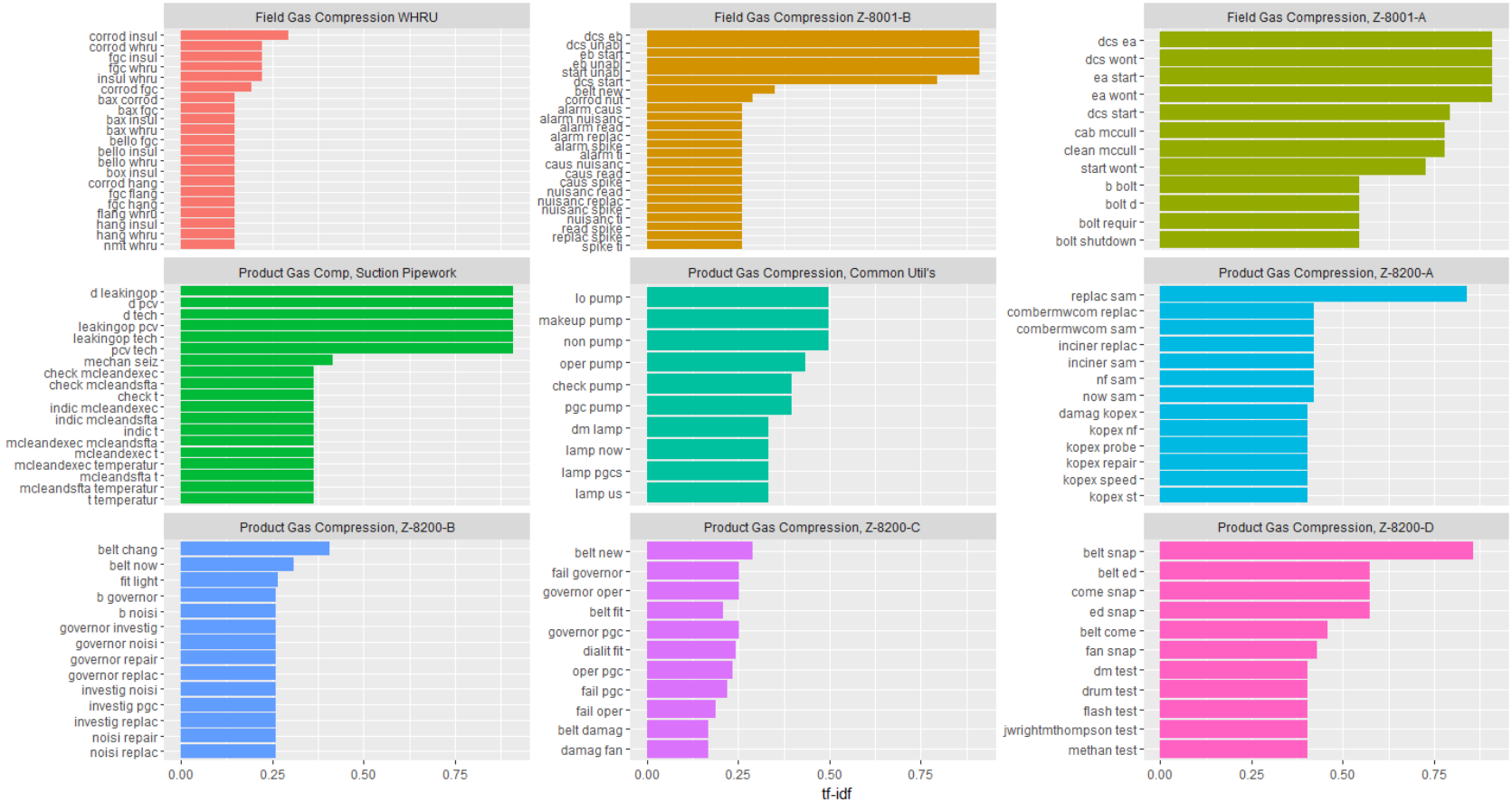
- Compute tf-idf for all terms:





Information Retrieval: tf-idf scheme (4)

centrica



Measuring Correlation (1)

centrica

D-8001-A/B Corroded Valve Bonnet Bolts

25·07·2003 04:09:24 MICHAEL WHITE (WHITEM), D-8001-A/B Corroded Valve Bonnet Bolts, Bolts fastening down valve bonnets corroded away;, D-8001-A LP Stage - V-80021, V-80522, V-80032, D-8001-B HP Stage - V-80521, V-80022, V-80532

Ref: 74321346542

{dab, corrod, valv, bonnet, bolt, fasten, da, lp ...}

{da, corrod, bolt, valv, hand, field, gas...}

D-8001-A Corroded Bolts & Valve Handle

29·04·2004 07:37:46 Mike White (WHITEM9), D-8001-A Corroded Bolts Valve Handle, Field Gas Suction Discharge Train A - D-8001-A, Please see attached document for issues;, Page 1 - Photo 20 = Corroded nuts and studs require replacement·

Ref: 457683143456

Measuring Correlation (2)

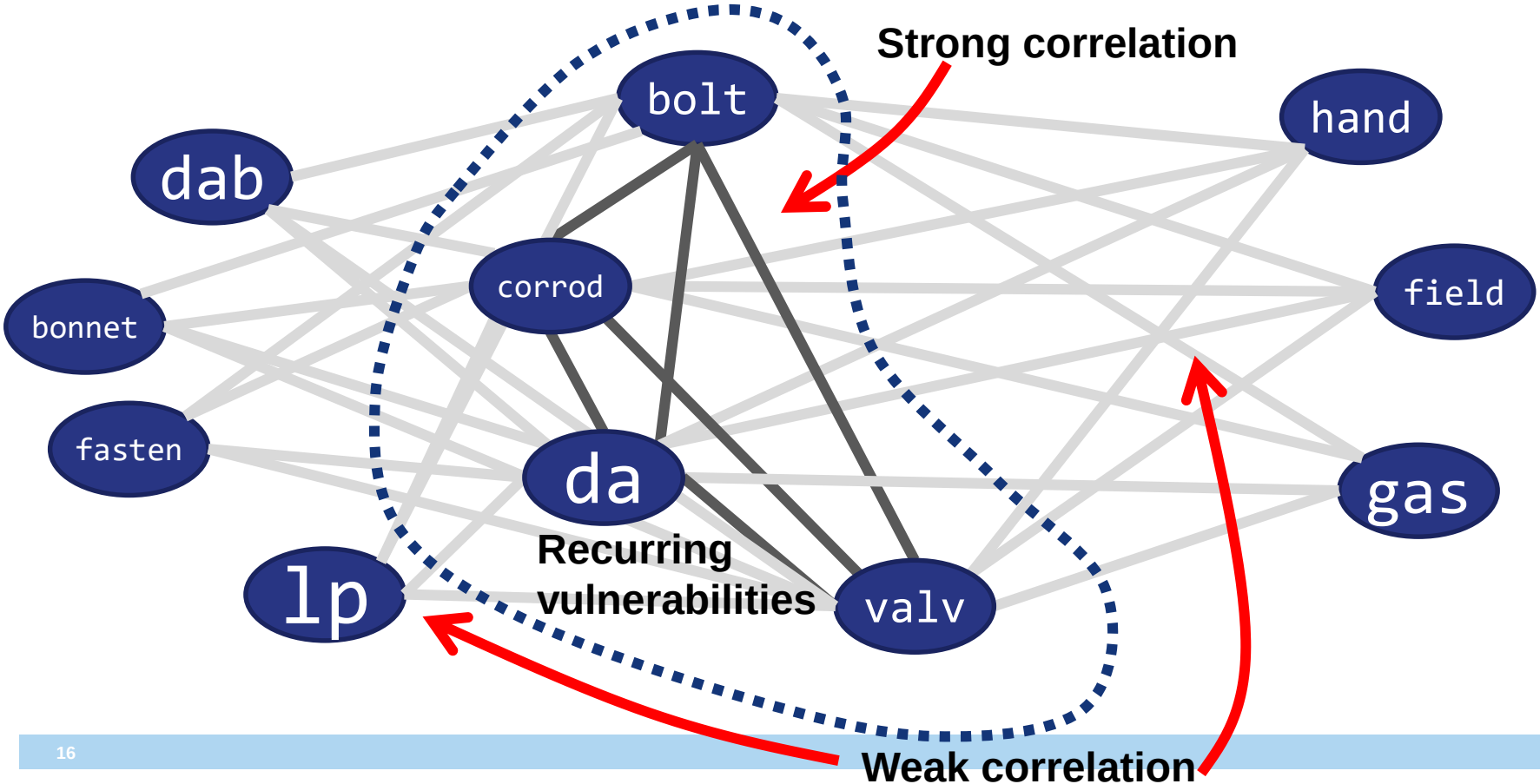
centrica

- Compute pairwise correlation for all terms

	dab	corrod	valv	bonnet	bolt	fasten	da	lp	...
dab									
corrod	0.11								
valv	0.03	0.24							
bonnet	0.05	0.32	0.27						
bolt	0.03	0.39	0.24	0.16					
fasten	0.15	0.13	0.26	0.17	0.35				
da	0.23	0.09	0.16	0.11	0.13	0.10			
lp	0.21	0.12	0.14	0.09	0.12	0.13	0.15		
...	...	...	...	...	...	...	...	...	

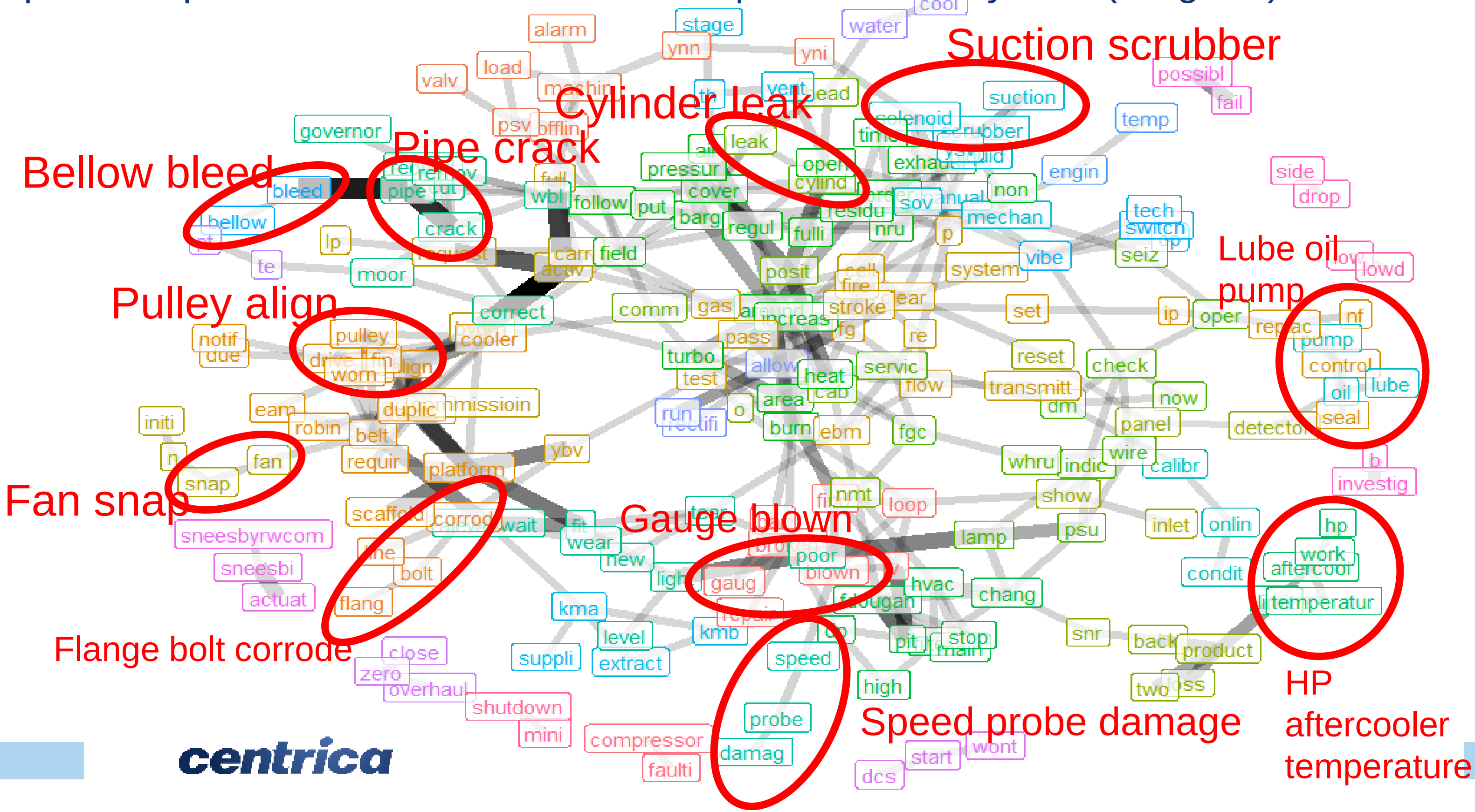
Measuring Correlation (3)

centrica



Identifying Vulnerability (1)

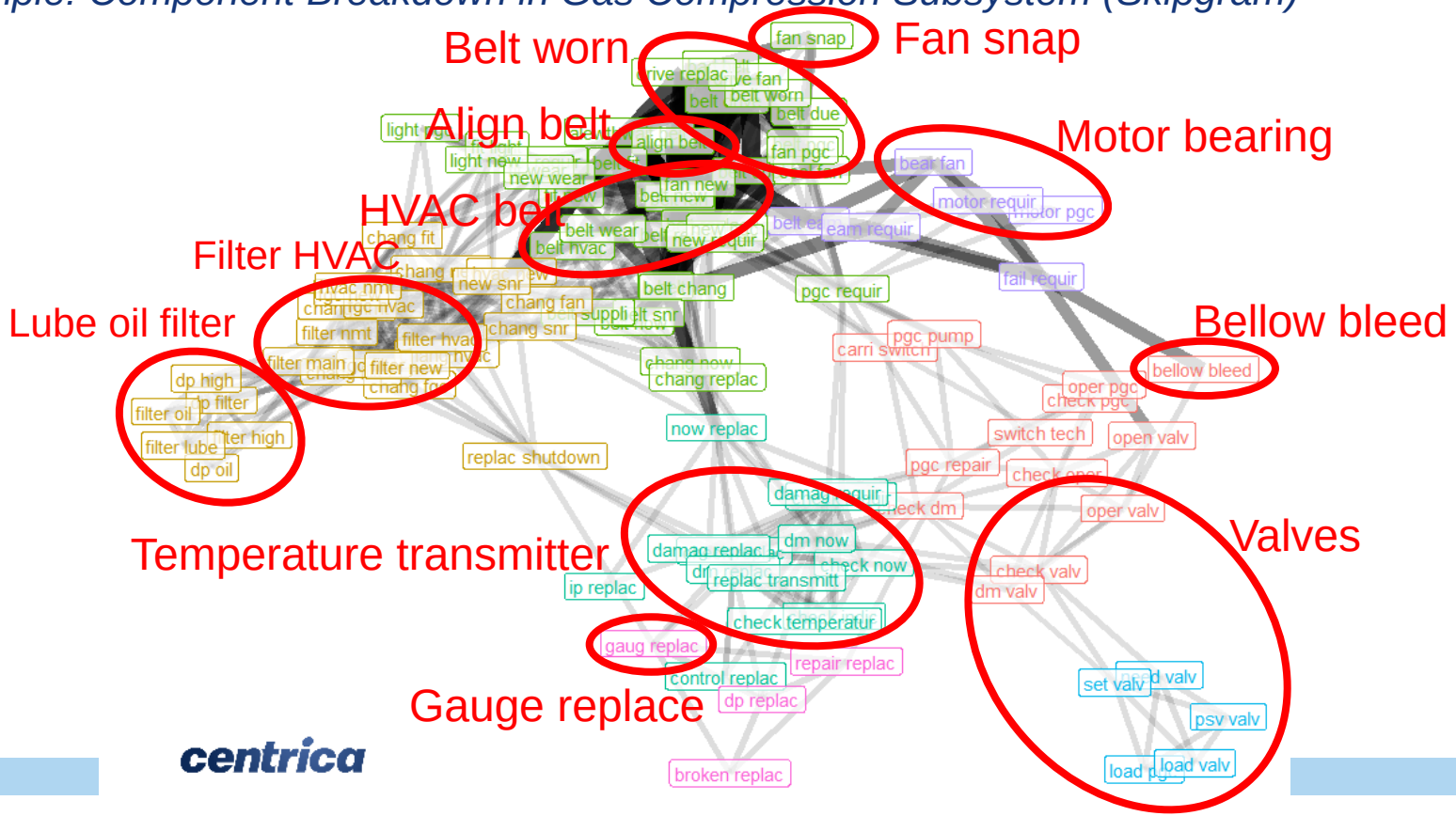
Example: Component Breakdown in Gas Compression Subsystem (Unigram)



Identifying Vulnerability (2)

centrica

Example: Component Breakdown in Gas Compression Subsystem (Skipgram)

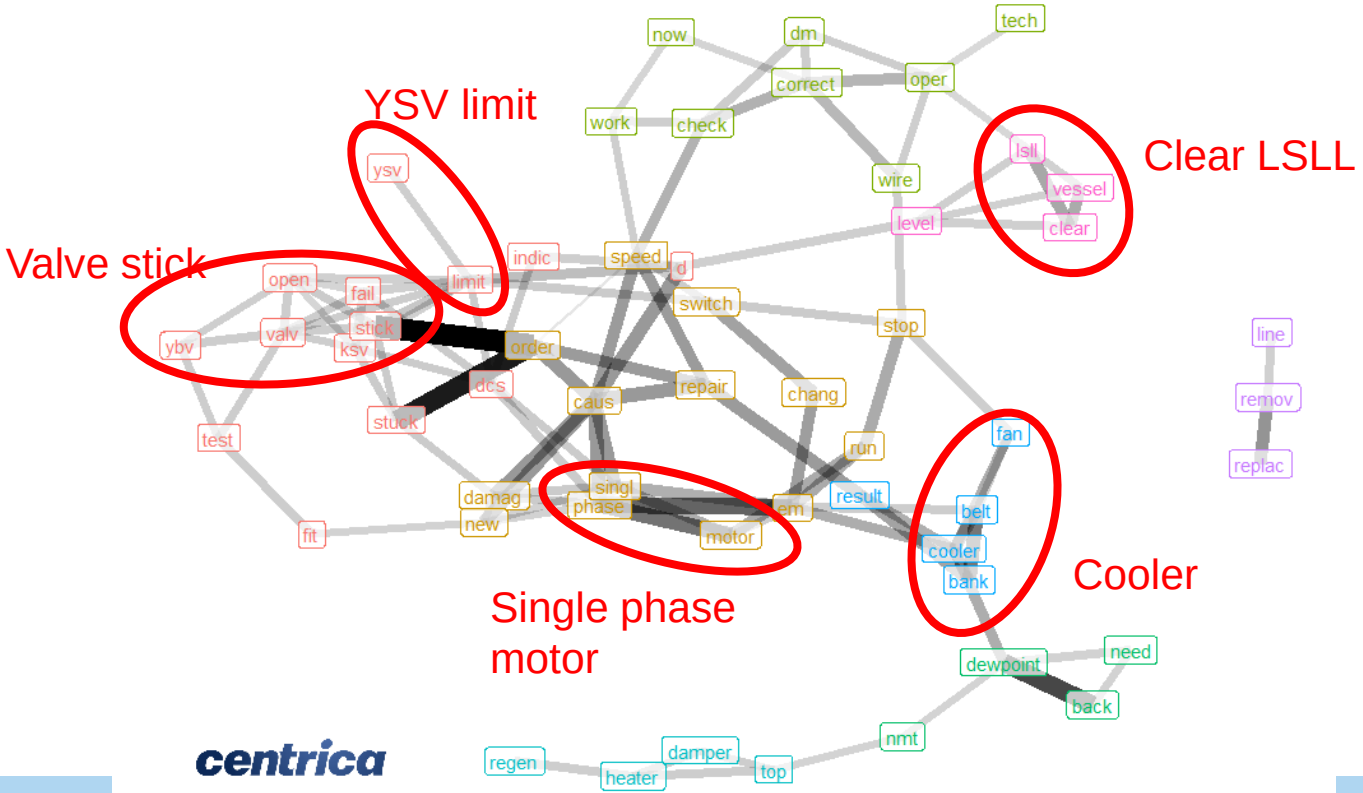


centrica

Identifying Vulnerability (3)

centrica

Example: Component Breakdown in Dewpoint Control Subsystem (Unigram)

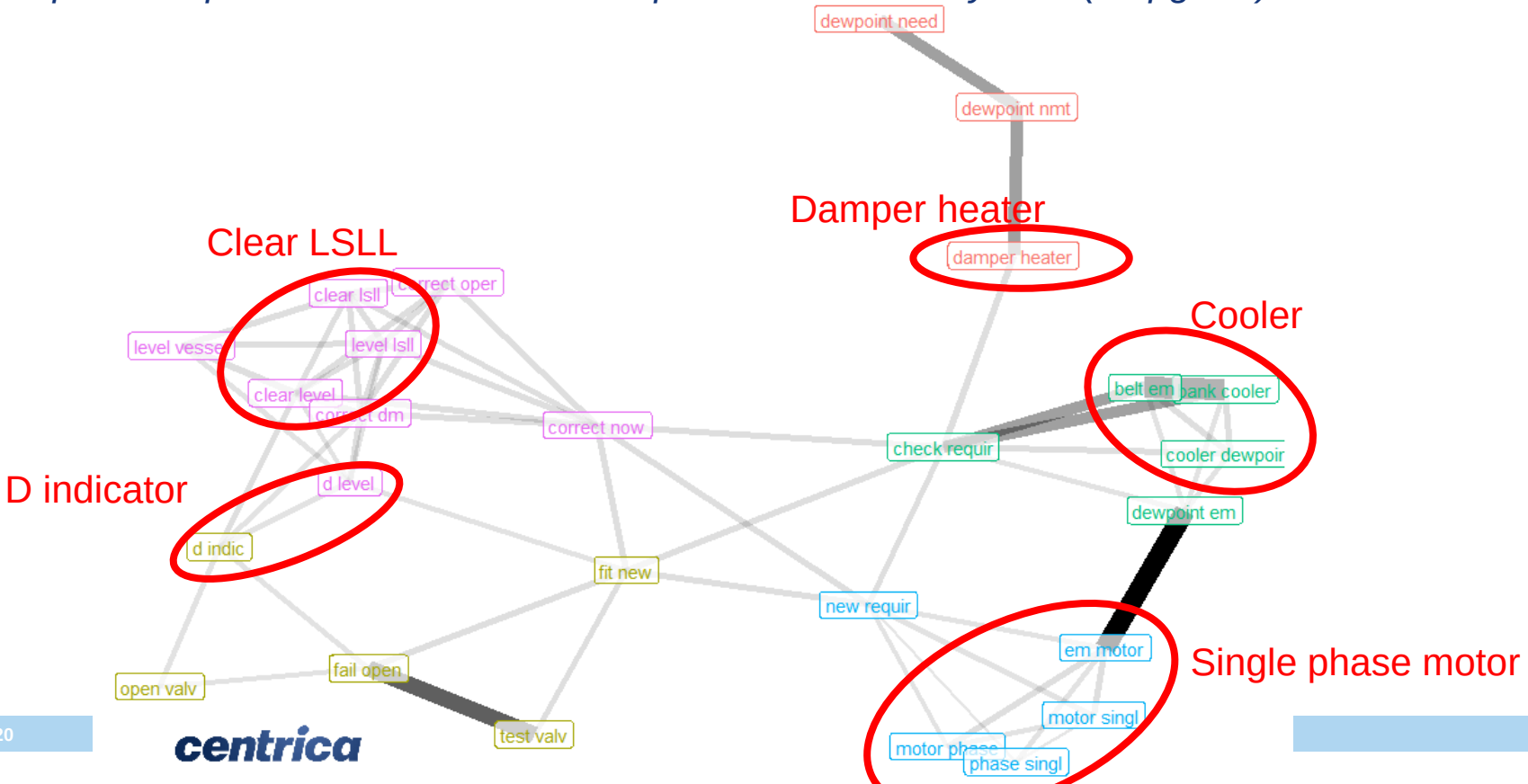


centrica

Identifying Vulnerability (4)

centrica

Example: Component Breakdown in Dewpoint Control Subsystem (Skipgram)



centrica

Identifying Vulnerability (5)

centrica

- Brainstorming workshop
  - Technical expertise from Process Engineer
  - Highlighted recurring issues manually
  - Close up investigation

Q&A

centrica

Timothy Wong
Senior Data Scientist, Centrica plc
timothy.wong@hotmail.co.uk