SAND2011-3485C

Text Analysis and Social Simulation

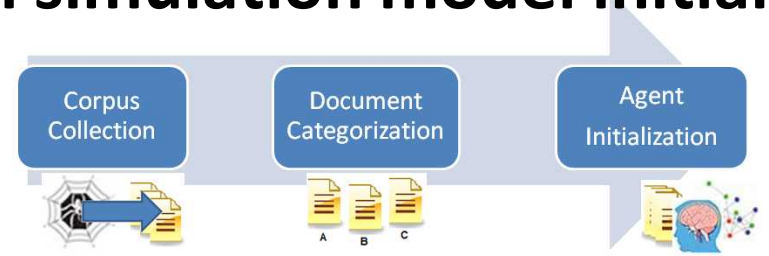
Zach Heath zheath@sandia.gov
JT McClain jtmccl@sandia.gov

Sandia National Laboratories

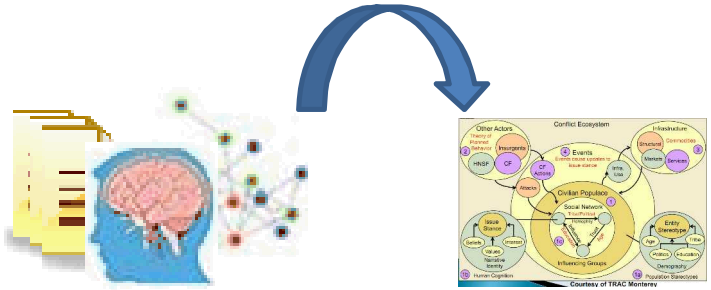
Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

TRAC-MTRY/Sandia National Laboratories Collaboration

- Two Main Tasks
  - Create a **modular pipeline** that uses **text analysis** for **social simulation model initialization**



  - Augment TRAC's Cultural Geography (CG) Model with **cognitive agents** derived from **text analysis**



Motivation

- Agent based model initialization is difficult
  - Manual source material collection and subject matter expertise is expensive
  - Subject matter expertise injects subjectivity

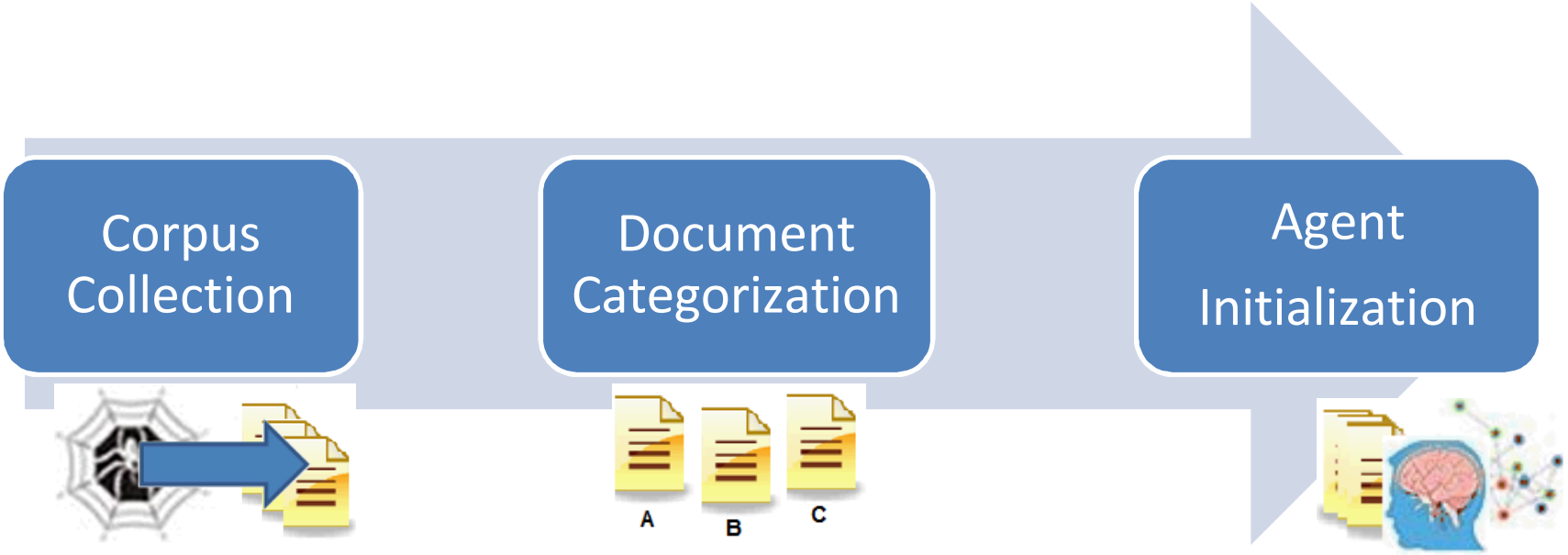
- Text analysis can **automate**, **augment**, and provide **objectivity** to much of this process

Caveats

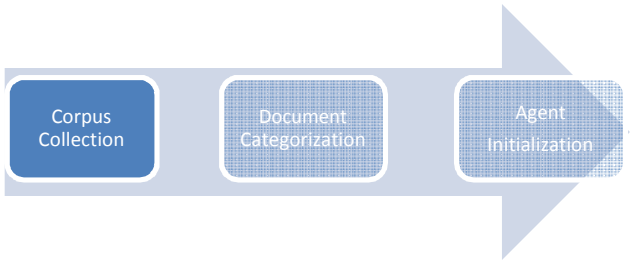
- This is in the proof of concept stage.
- This has not been validated
- Nowhere near being able to remove subject matter expertise from the process

Agent Initialization Through Text Analysis

- Task 1 Goals
  - Create modular pipeline for agent initialization process
  - Provide initial test implementations for each module
  - Run and measure pipeline implementation using a test case



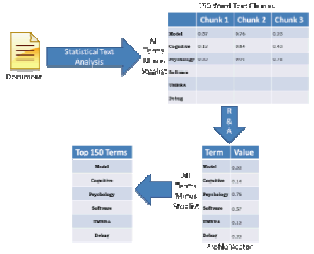
Corpus Collection



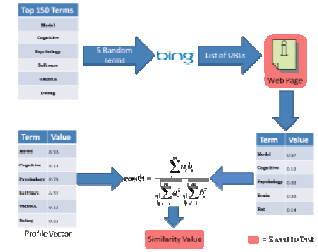
- Purpose:
  - Automatically or manually collect a set of documents that will be analyzed for agent based model initialization
- Inputs:
  - Example source document of interest and/or search keywords
  - Location to search
    - Directory or Database of Files
    - List of web domains
    - The internet as a whole
- Outputs:
  - Directory of documents
  - List of documents with model matching score

Basic Crawler Process

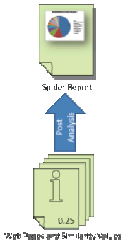
Document



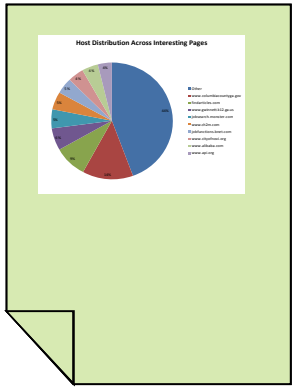
[Document Preprocessing]



[Document Collection]



[Post Analysis]



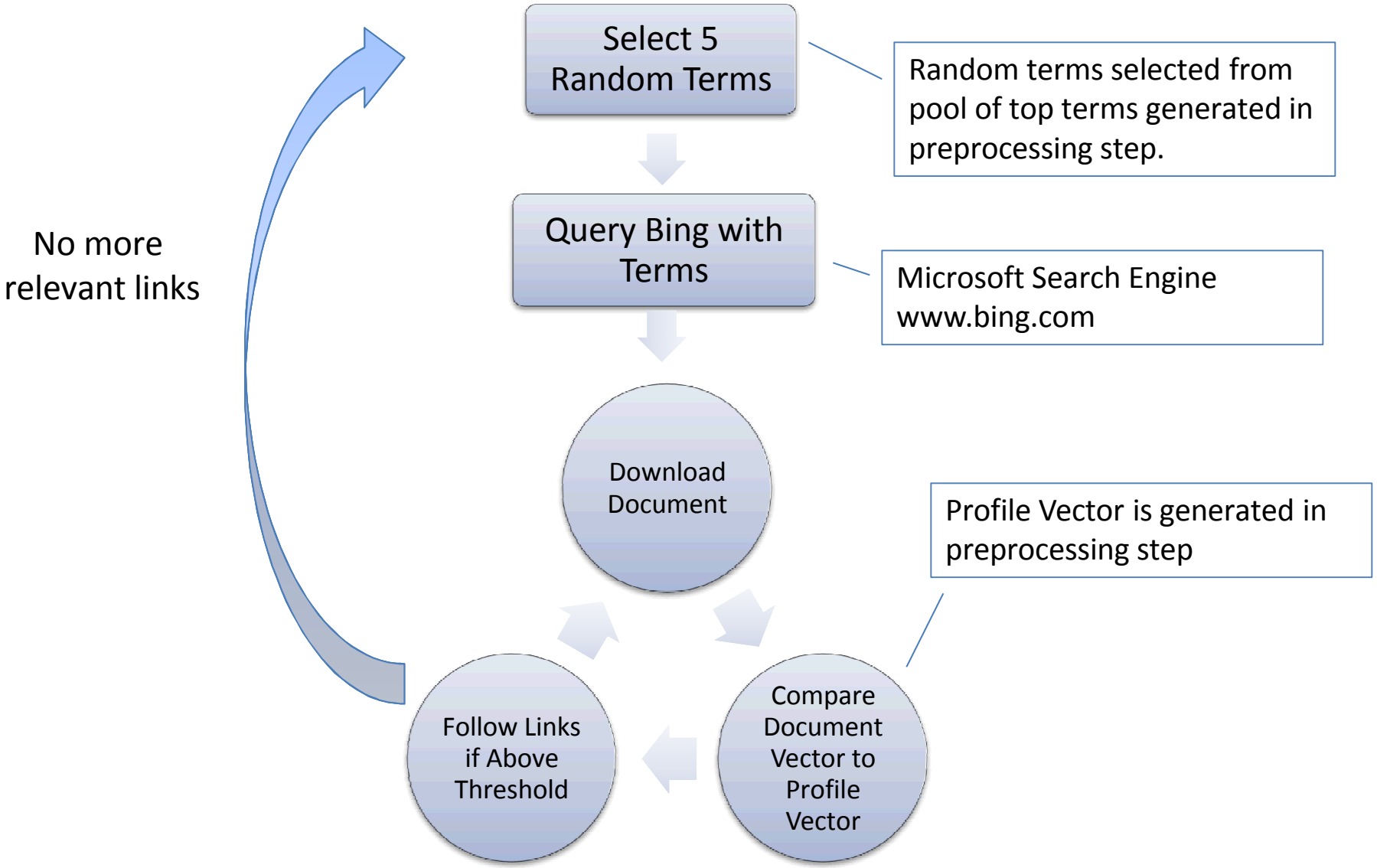
Crawler Report

Analyze seed documents for top terms
Create text profile of seed documents

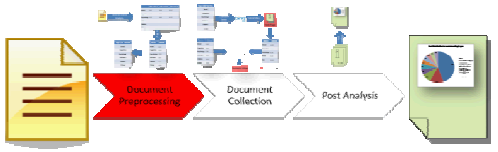
Gather documents that "match" the text profile

Aggregated list of documents, matching scores, and link summaries

Basic Crawling Process



Document Preprocessing



Document

Statistical Text Analysis

All Terms Minus Stoplist

150 Word Text Chunks

	Chunk 1	Chunk 2	Chunk 3
Model	0.57	0.76	0.23
Cognitive	0.12	0.84	0.43
Psychology	0.32	0.91	0.75
Software			
UMBRA			
Debug			

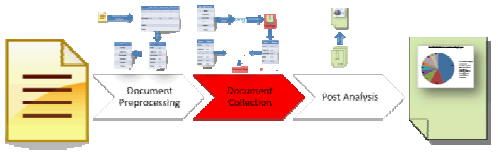
Profile Vector

Term	Value
Model	0.23
Cognitive	0.14
Psychology	0.76
Software	0.57
UMBRA	0.12
Debug	0.32

All Terms Minus Stoplist

Top 150 Terms
Model
Cognitive
Psychology
Software
UMBRA
Debug

Document Collection



Top 150 Terms
Model
Cognitive
Psychology
Software
UMBRA
Debug

5 Random Terms

bing

List of URLs

Web Page

Term	Value
Model	0.57
Cognitive	0.12
Psychology	0.32
Brain	0.23
Rat	0.14

Term	Value
Model	0.23
Cognitive	0.14
Psychology	0.76
Software	0.57
UMBRA	0.12
Debug	0.32

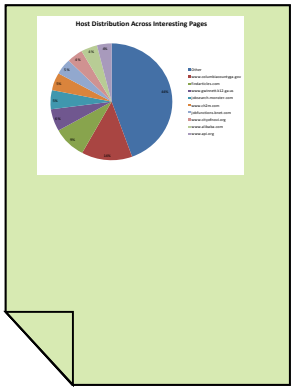
Profile Vector

$$\cos\Theta = \frac{\sum_{i=1}^{m} a_i b_i}{\sqrt{\sum_{i=1}^{m} a_i^2}\sqrt{\sum_{i=1}^{m} b_i^2}}$$

Similarity Value

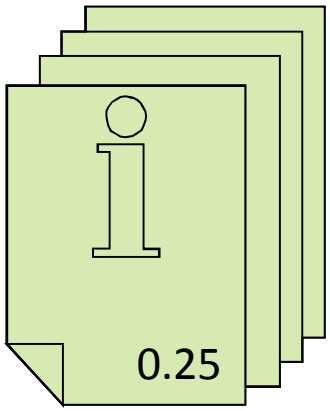
= Saved to Disk

Post Analysis



Crawler Report

Post Analysis



Web Pages and Similarity Values

Produces aggregated list of collected documents with scores indicating level of match to source documents.

Collected documents can be used to produce a new profile for the corpus

Retrieved documents can be accessed through Google like search interface

Crawler report provides pie charts that indicate ‘interesting’ web domains

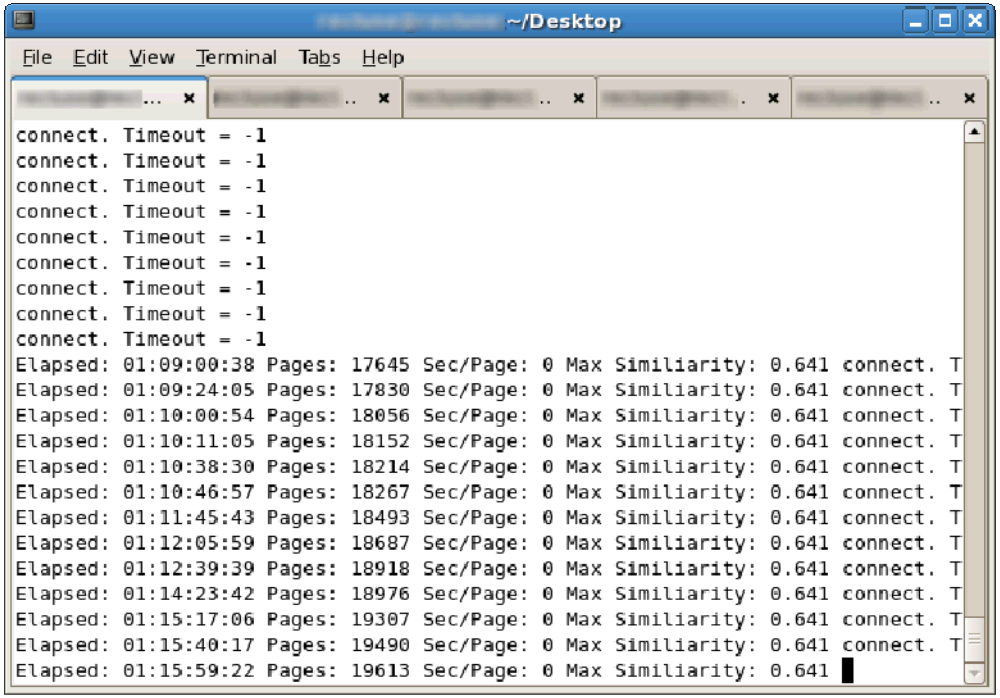
Cognitive Crawler

Seed document model is built from example documents

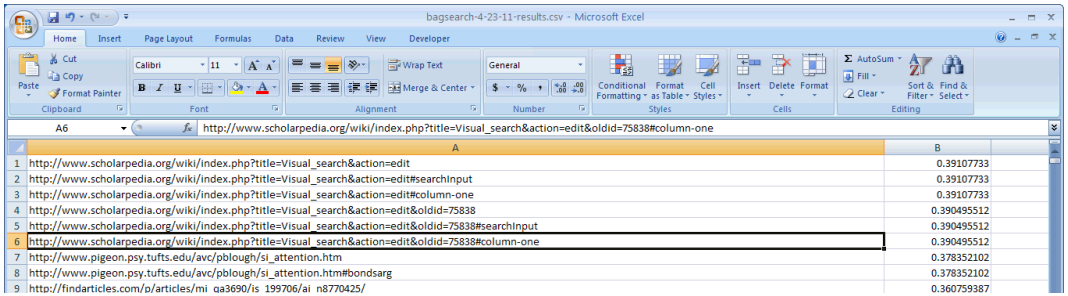
Cognitive Crawler is launched through command line. Console output tracks crawler progress

Crawler can run for weeks but generally finds most relevant material in first 1-2 days

Produces 15-20 GB of data



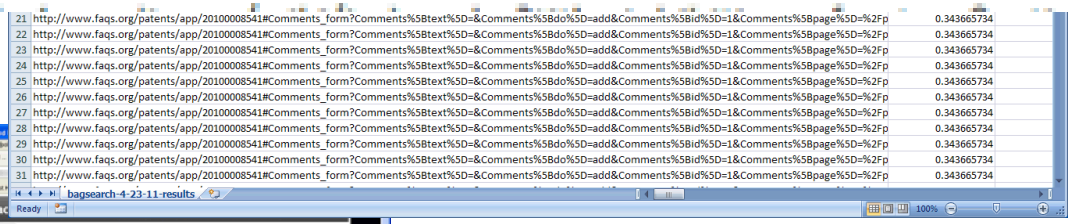
Cognitive Crawler



http://www.pigeon.psy.tufts.edu/avc/pblough/si_attention.htm#bondsarg

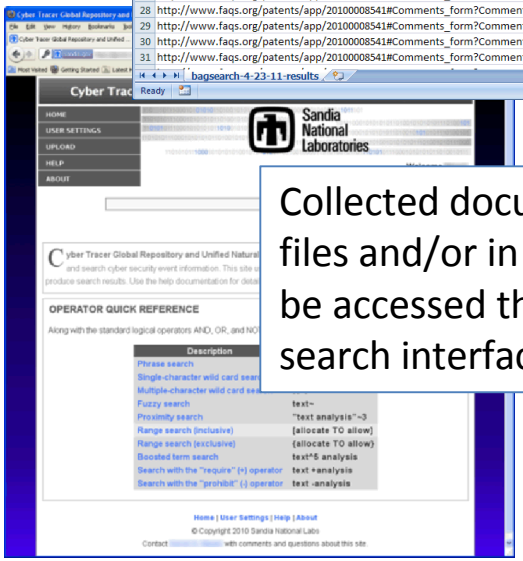
http://findarticles.com/p/articles/mi_qa3690/is_199706/ai_n8770425/

http://findarticles.com/p/articles/mi_qa3690/is_199706/ai_n8770425/#talkback

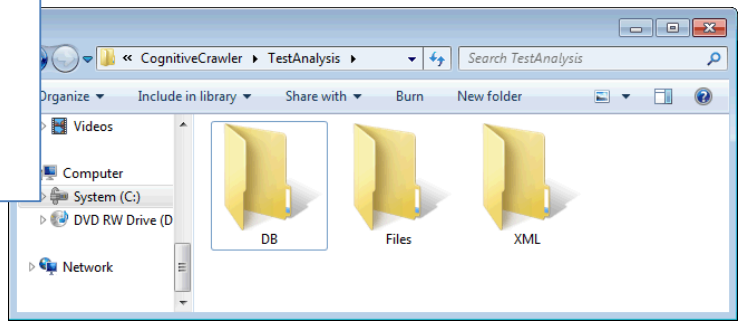


Score represents how well collected documents matches text profile generated in preprocessing step

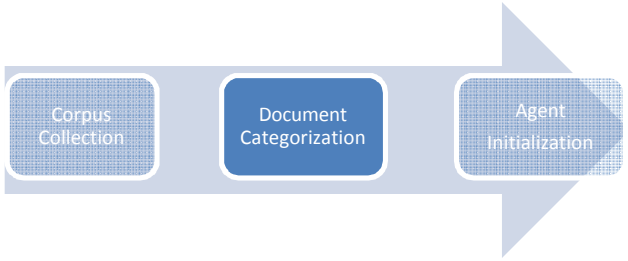
0.378352102
0.360759387
0.356080931
0.343753779
0.343753779
0.343753779
0.343665734



Collected documents stored as flat files and/or in a database. Files can be accessed through Google like search interface

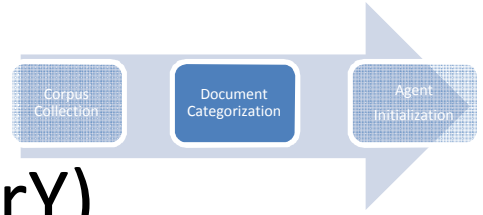


Document Categorization



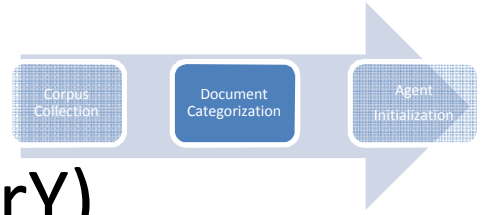
- Purpose:
  - Given a collection of documents, group or categorize those documents based on some set criteria
- Inputs:
  - Document Corpus
  - Categorization criteria and pre-categorized documents
  - Categorization thresholds
- Outputs:
  - Document Categorizations
  - Categorization Statistics

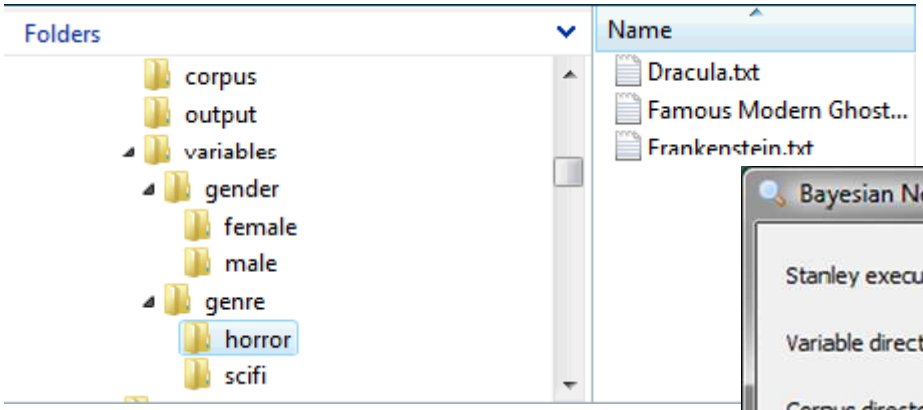
Proof of concept: STANLEY
(Sandia Text ANaLysis Extensible librarY)
Document Categorizer

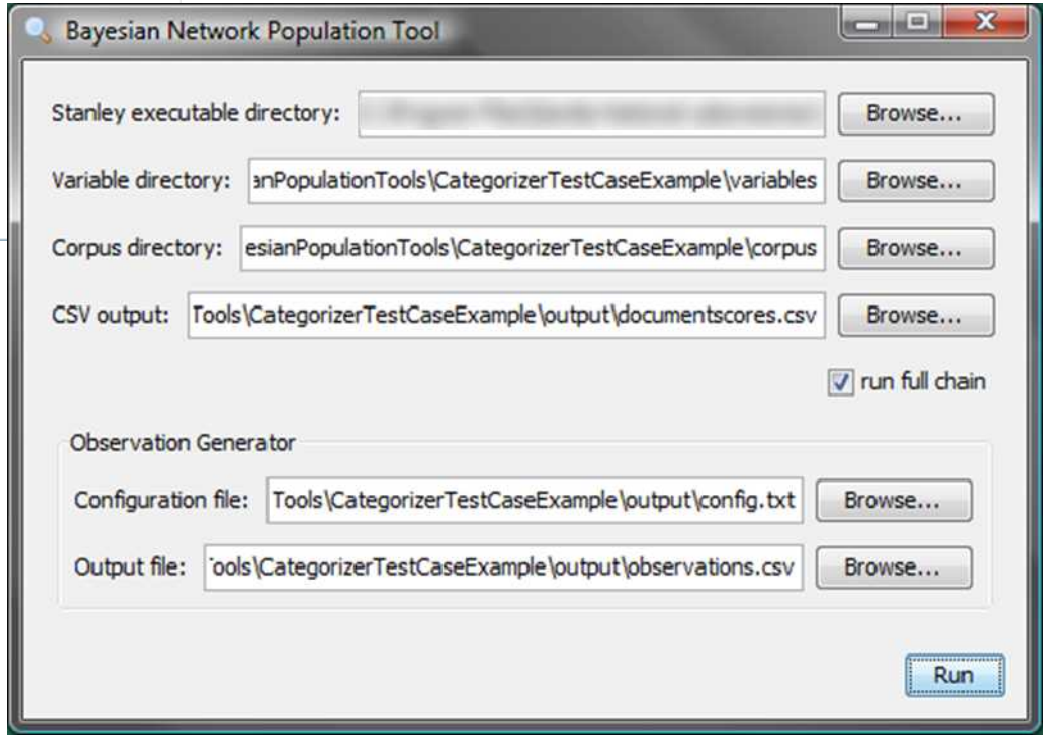


- Builds textual profile for each set of pre-categorized documents
- Matches and scores new documents from corpus against all profiles using same text analysis tools explained in Cognitive Crawler section
- Marks document categorizations based on thresholds set by user
- To be replaced with Cortext (java based text analysis library)

Proof of concept: STANLEY
(Sandia Text ANaLysis Extensible librarY)
Document Categorizer

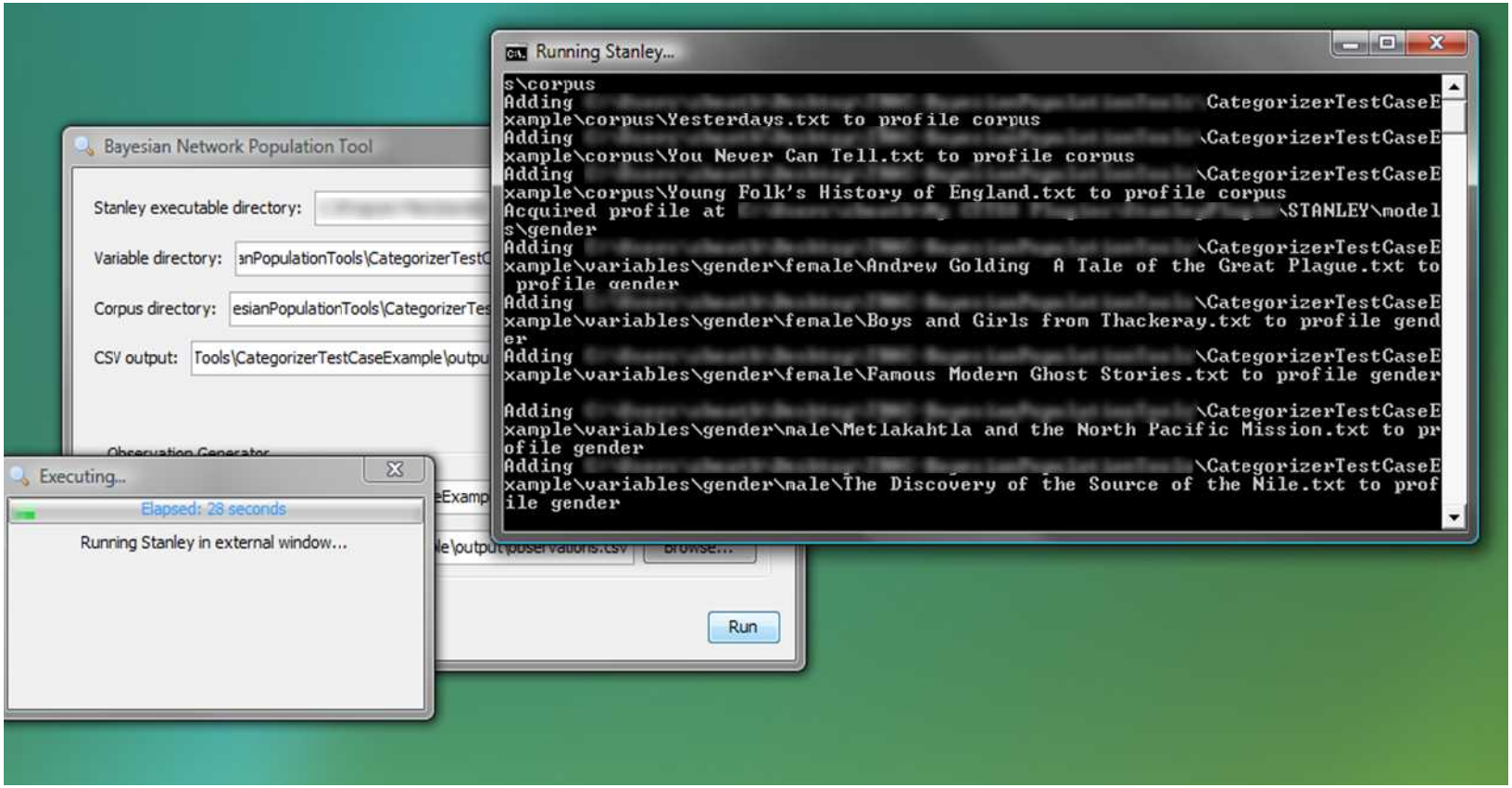




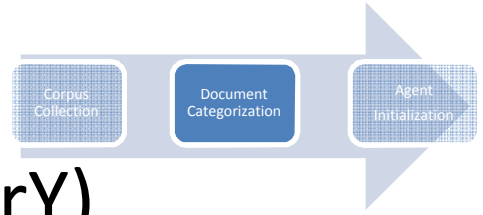


Directory structure represents variables to categorize on. User must fill these directories with example documents

Proof of concept: STANLEY
(Sandia Text ANaLysis Extensible librarY)
Document Categorizer



Proof of concept: STANLEY
(Sandia Text ANaLysis Extensible librarY)
Document Categorizer



documentScores.csv - Microsoft Excel

13	@Data					
14	95 Theses.txt	corpus	0.51786308			
15	95 Theses.txt	authorGender	0.531130992	0.525042539	0.541182	
16	95 Theses.txt	genre	0.45101695	0.619124029	0.292756	0.525762
17	95 Theses.txt	timeWritten	0.583643893	0.562309539	0.548094	0.596231
18	95 Theses.txt	violence	0.57429249	0.571978532	0.548111	

How well document matches the example corpus as a whole

How well document matches violence variable 1 and 2 example documents

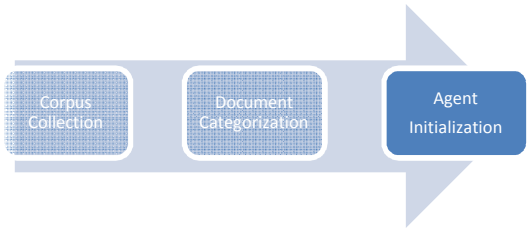
How well document matches all violence variable example documents

	A					
1	% Document Score File					
2	% Corpus Documents:					
3	% Variable Documents:					
4	% Date: 10/29/2010 11:57 AM					
5						
6	@Variables					
7	corpus					
8	authorGender	female	male			
9	genre	biography	cooking	horror		
10	timeWritten	1700sAndEarlier	1800s	1900s		
11	violence	acceptable	unacceptable			
12						
13	@Data					
14	95 Theses.txt	corpus	0.51786308			
15	95 Theses.txt	authorGender	0.531130992	0.525042539	0.541182	
16	95 Theses.txt	genre	0.45101695	0.619124029	0.292756	0.525762
17	95 Theses.txt	timeWritten	0.583643893	0.562309539	0.548094	0.596231
18	95 Theses.txt	violence	0.57429249	0.571978532	0.548111	
19	A Confederate Girl's Diary.txt	corpus	0.923951505			
20	A Confederate Girl's Diary.txt	authorGender	0.936586042	0.94070377	0.923283	
21	A Confederate Girl's Diary.txt	genre	0.87281851	0.927071982	0.816778	0.927507
22	A Confederate Girl's Diary.txt	timeWritten	0.942875824	0.890148662	0.930929	0.93014
23	A Confederate Girl's Diary.txt	violence	0.961517021	0.956108888	0.948304	
24	A Connecticut Yankee in King Arthur's Court.txt	corpus	0.918447855			
25	A Connecticut Yankee in King Arthur's Court.txt	authorGender	0.933656124	0.935281377	0.923245	

Agent Initialization

- Purpose:
  - Analyze a set of document categorizations or other text analysis operations to produce data for agent based model initialization
- Inputs:
  - Document Categorizations
  - Agent Initialization parameters
- Outputs:
  - Initialization files for Agent based model of interest
  - Additional Statistics

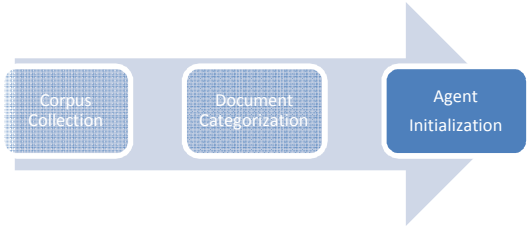
Proof of concept: Observation Generator



- Analyzes document categorization scores to produce list of document observations

- Document observations are turned into a Bayesian Network for use in the Cultural Geography Model

Proof of concept: Observation Generator



Categorization scores from previous step

documentScores.csv - Microsoft Excel

	A	B	C	D
1	% Document Score File			
2	% Corpus Documents:	\corpus		
3	% Variable Documents:	\variables		
4	% Date: 10/29/2010 11:57 AM			
5				
6	@Variables			
7	corpus			
8	authorGender	female	male	
9	genre	biography	cooking	horror
10	timeWritten	1700sAndEarlier	1800s	1900s
11	violence	acceptable	unacceptable	
12				
13	@Data			
14	95 Theses.txt	corpus		
15	95 Theses.txt	authorGender		
16	95 Theses.txt	genre		
17	95 Theses.txt	timeWritten		
18	95 Theses.txt	violence		
19	A Confederate Girl's Diary.txt	corpus		
20	A Confederate Girl's Diary.txt	authorGender		
21	A Confederate Girl's Diary.txt	genre		
22	A Confederate Girl's Diary.txt	timeWritten		
23	A Confederate Girl's Diary.txt	violence		
24	A Connecticut Yankee in King Arthur's Court.txt	corpus		

Observation Generator

Setup | Input | Config | Ouput

Input file: ools\CategorizerTestCaseExample\output\documentscores.csv Browse...

Configuration file: nTools\CategorizerTestCaseExample\output\config.txt Browse...

Output file: \Tools\CategorizerTestCaseExample\output\observations.csv Browse...

Generate ARFF output file

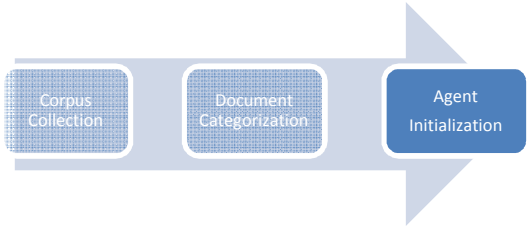
cfBlais9.txt - Notepad

File Edit Format View Help

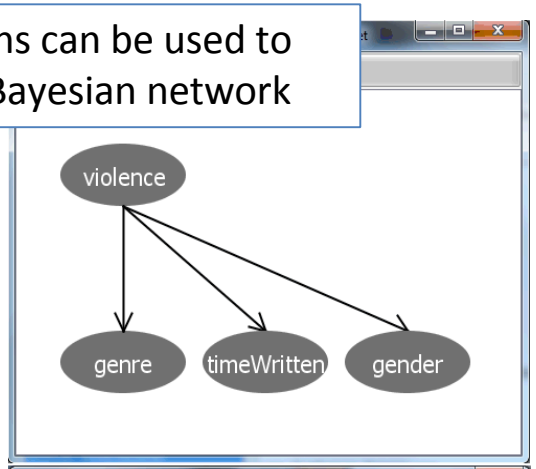
```
# Configuration Threshold file for converting document scores to ob
# Format: VAR_NAME, VAR_OPTION_NAME, VAR_SCORE, VAR_OPTION_SCORE, \
#    VAR_NAME: Name of the variable to be configured (ex:  genre)
#    VAR_OPTION_NAME : The name of the variable output value to be
#    VAR_SCORE: The minimum score value that a document must have f
#    VAR_OPTION_SCORE: The minimum score value that a document must
#    VAR_OPTION_DIFF:  The minimum score difference that the docume
# "*" matches any variable name or any variable output value
# *,*,.6,.6,.01 Sets the threshold for all variables and all variab
# genre,*,.6,.6,.01 Sets the threshold for all variable options of
# genre,fiction,.6,.6,.01 Sets the threshold for the variable optic
# Rules are matched by the order they are specified within this fi

violence, unacceptable,.3,.3,.001
violence, acceptable,.8,.85,.05


*,*,.3,.3,.005
```

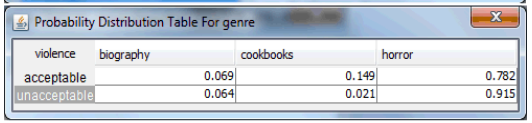
Thresholds for matching to be set by user

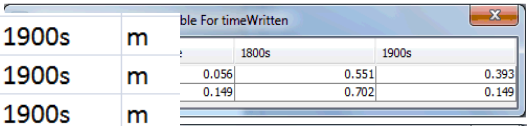
Proof of concept: Observation Generator

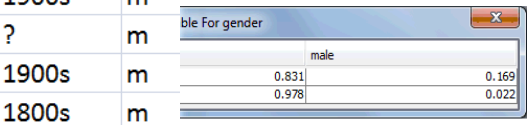


Observations can be used to initialize a Bayesian network

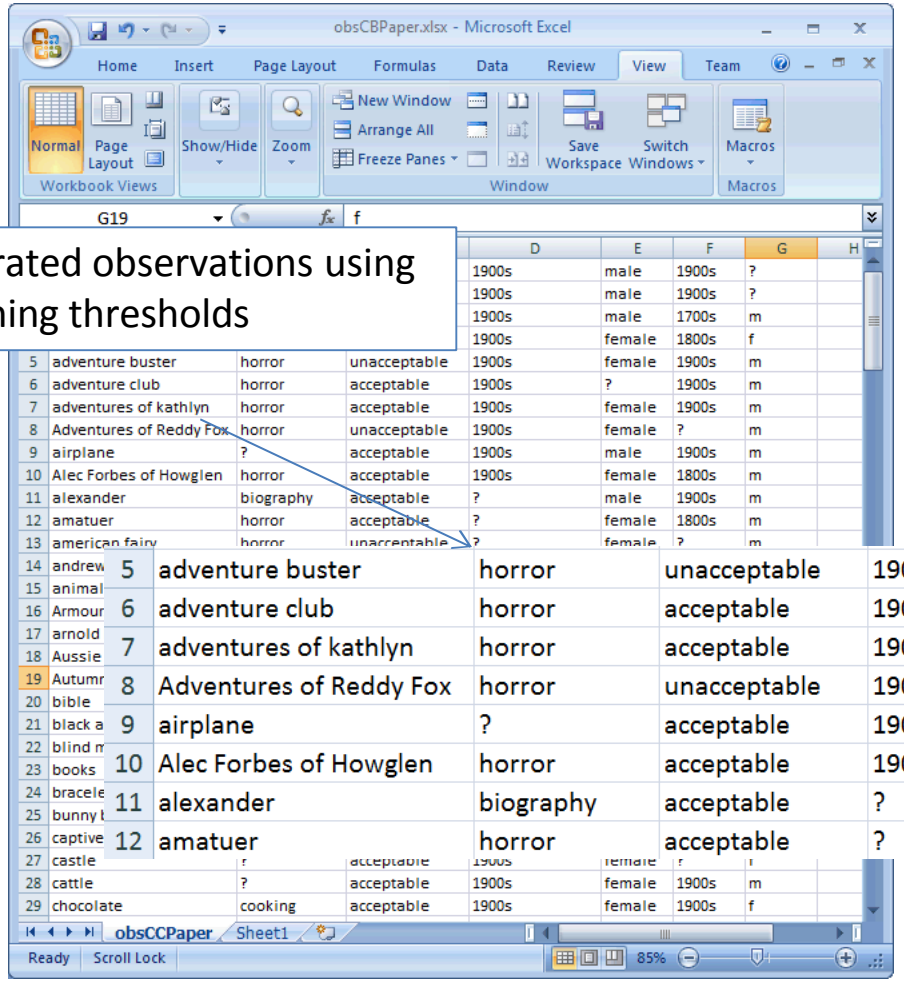








Generated observations using matching thresholds



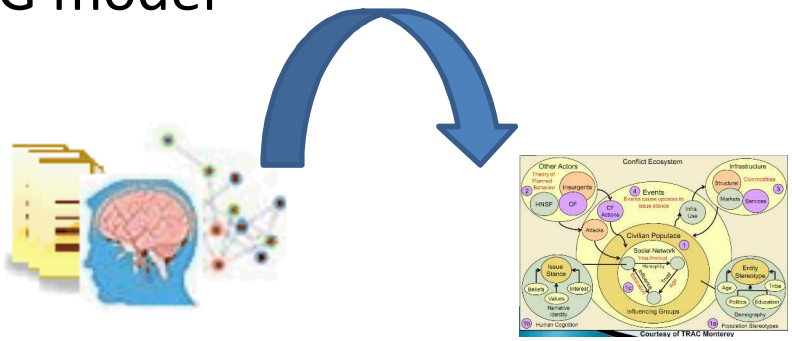
5	adventure buster	horror	unacceptable	1900s	female	1900s	m
6	adventure club	horror	acceptable	1900s	?	1900s	m
7	adventures of kathlyn	horror	acceptable	1900s	female	1900s	m
8	Adventures of Reddy Fox	horror	unacceptable	1900s	female	?	m
9	airplane	?	acceptable	1900s	male	1900s	m
10	Alec Forbes of Howglen	horror	acceptable	1900s	female	1800s	m
11	alexander	biography	acceptable	?	male	1900s	m
12	amatuer	horror	acceptable	?	female	1800s	m

Agent Initialization Through Text Analysis

- Proof of concept implementations requiring subject matter expert guidance/analysis
- Pipeline will be modular to allow for testing of alternative/improved implementations
- Continual improvement of components will relieve burden on subject matter expert

Integration of Cognitive Model based Agents in CG

- Task 2 Goals
  - Integrate concept map based cognitive model agents into the CG model



  - Run and measure implementation using a test case comparing with other agent implementations

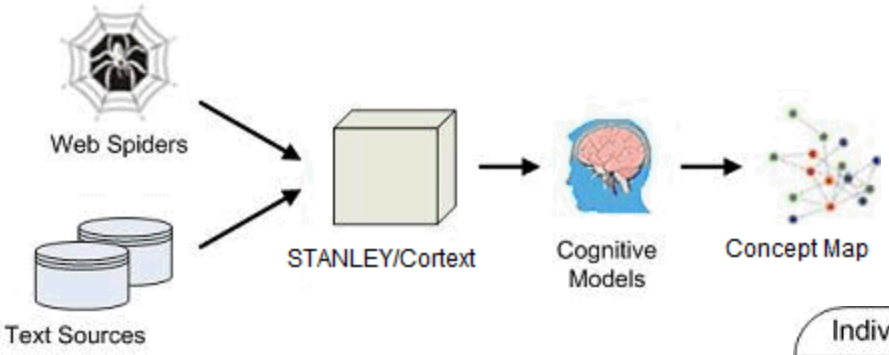
Concept map agents

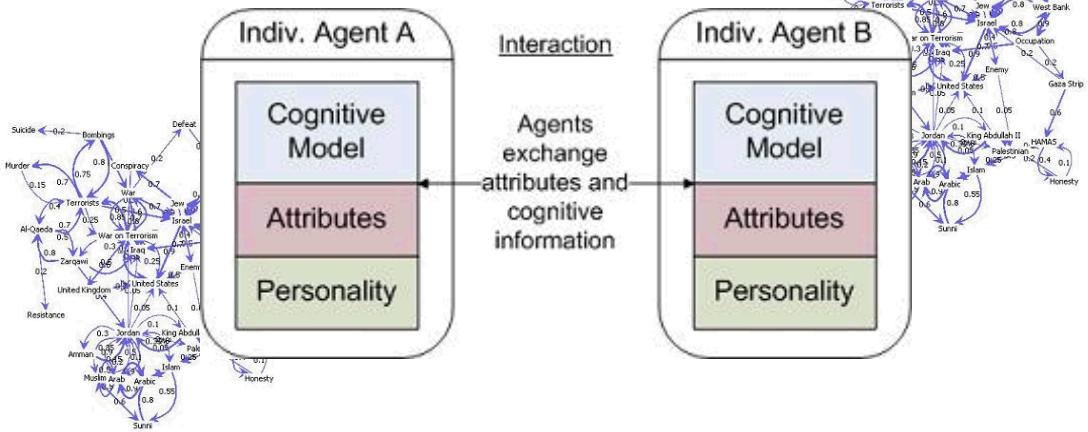
Individuals have cognitive models created from text

Cognitive information is exchanged during interactions

Media Agents also inject cognitive

information to their subscribers





Concept map agents