CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

http://www.cit.ie

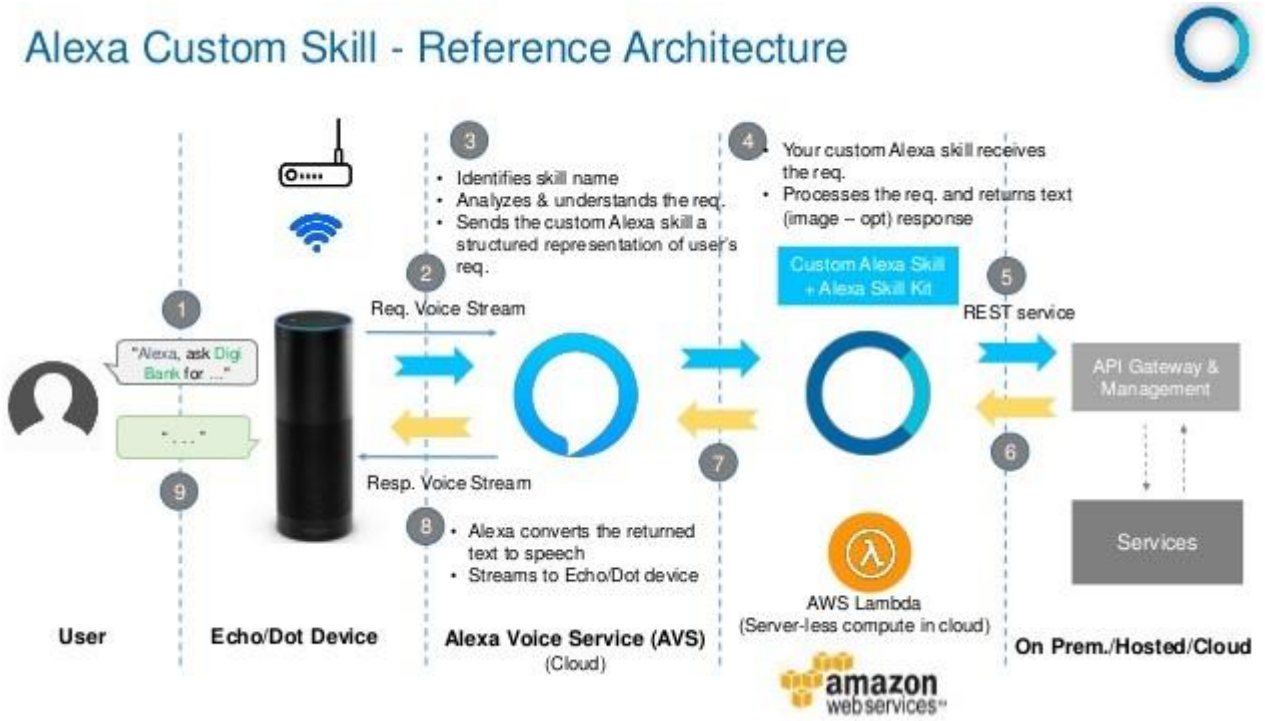
Natural Language Processing Lab

Week3: UNIX Lab

Praveen Joshi

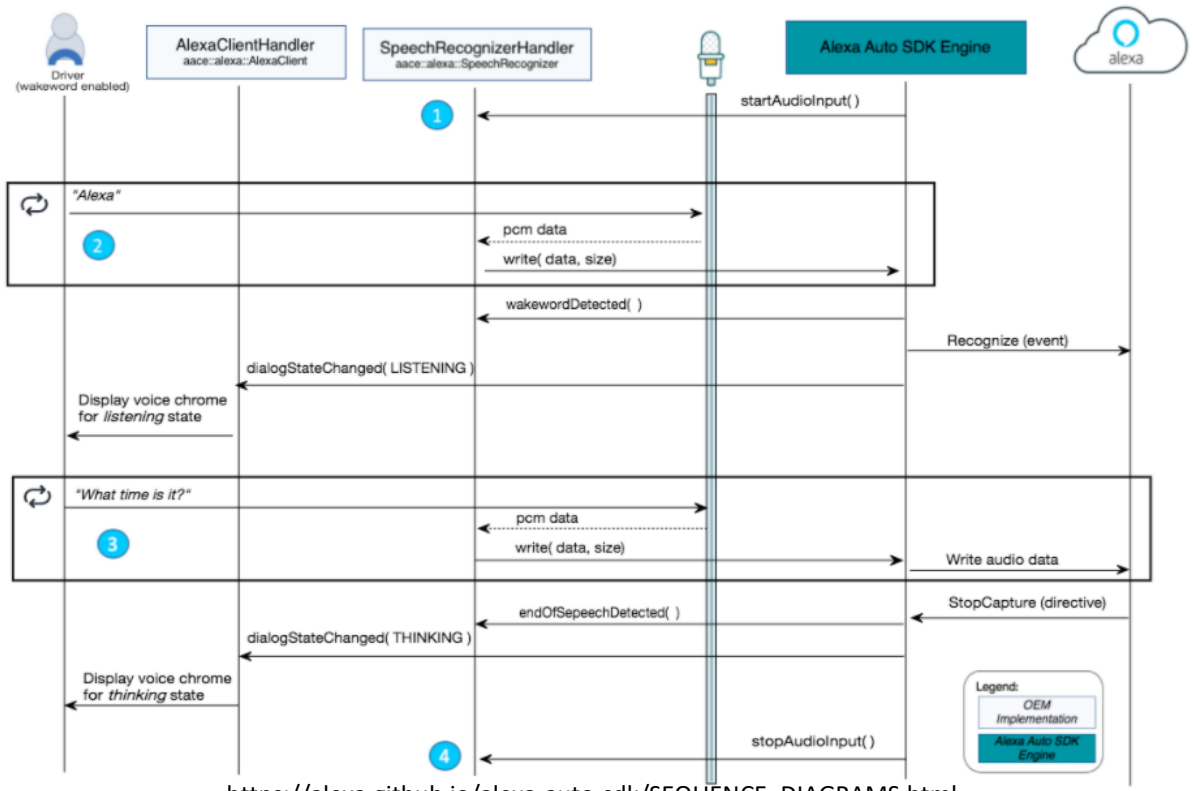
05/10/2020

Alexa – E2E



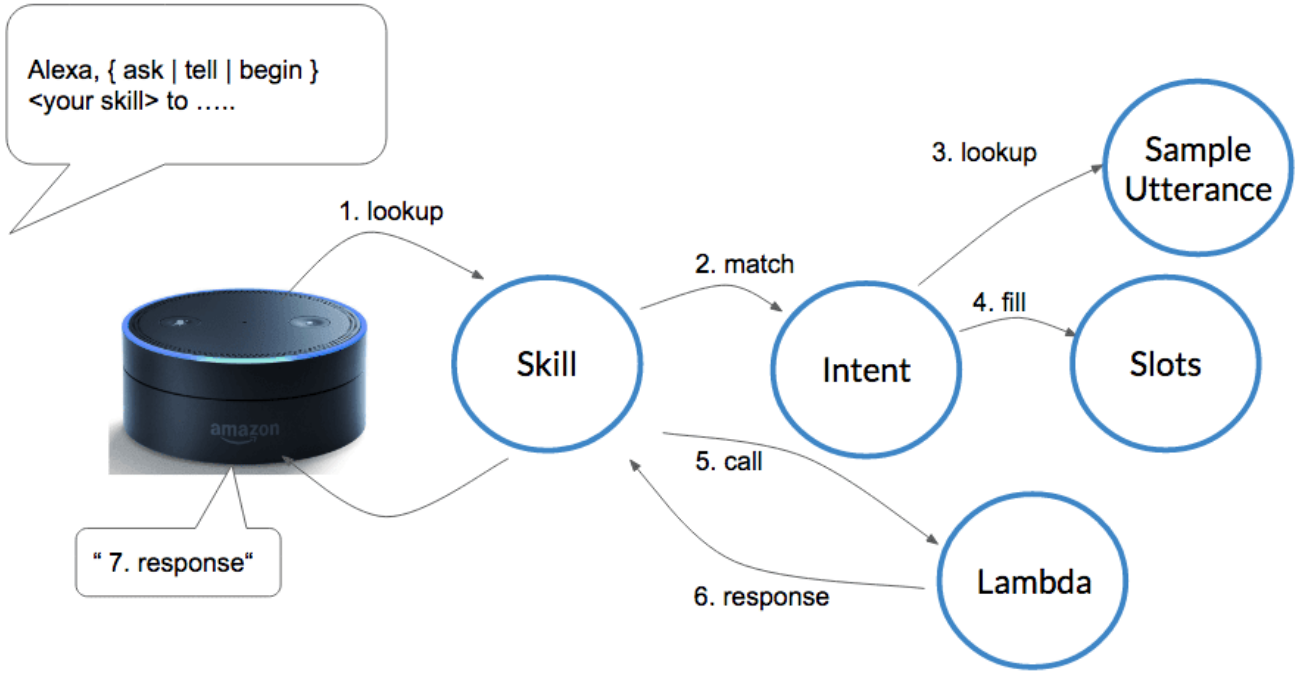
https://www.krminc.com/alexa-app-develpment/

Alexa - Engineering



https://alexa.github.io/alexa-auto-sdk/SEQUENCE_DIAGRAMS.html

Alexa - NLP



https://www.dipockdas.com/2016/11/create-alexa-skill-for-your-amazon.html

Regular Expressions

- Why?
  - Matching/Finding
  - Doing something with matched text
  - Validation of data
  - Case insensitive matching
  - Parsing data (ex: html)
  - Converting data into diff for etc.

Regular Expressions

- Entities to extract:
  - Board of Directors
  - Price at NAV
  - Fund Assets (millions)
  - Expense Ratio
  - Minimum Initial Investment
  - Fund Name, Class and series extraction
  - Annual Management Fees
  - Expense Ratio

ATLANTE FUNDS PLC

DIRECTORS
Peter Blessing
Dermot Butler
Stuart Anthony Williams
Matteo Rigginello

REGISTERED OFFICE
70 Sir John Rogerson's Quay
Dublin 2
Ireland

SECRETARY
Matsack Trust Limited
70 Sir John Rogerson's Quay
Dublin 2

SPONSOR
Albemarle Asset Management Limited
28-29 Dover Street
London W1S 4NA
United Kingdom

INVESTMENT MANAGER
Albemarle Asset Management Limited
28-29 Dover Street
London W1S 4NA
United Kingdom

ADMINISTRATOR
Bank of Ireland Securities Services Limited
New Century House
Mayor Street Lower
IFSC
Dublin 1
Ireland

https://github.com/praveenjoshi01/Hedge_Fund_Information_Extraction

Issues : Regular Expressions

- The process we just went through was based on fixing two kinds of errors
  - Matching strings that we should not have matched (there, then, other)
    - False positives (Type I)
  - Not matching things that we should have matched (The)
    - False negatives (Type II)

Slide taken from Dr. Haithem Afli's Lecture on NLP

Confusion Matrix

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment) positives	Actual Value (as confirmed by experiment) negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

https://alearningaday.blog/2016/09/14/confusion-matrix/

Confusion Matrix

- Set of words: {the, The, There, then, other, there}
- Ground Truth: {the, The}
- Regex(Set of words): {the, there, then, other}

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment)	
		positives	negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

https://alearningaday.blog/2016/09/14/confusion-matrix/

	Actual Value		
Predicted Value		Positives	Negatives
	Positives	the	there, then, other
	Negatives	The	There

Confusion Matrix – Regex Expressions

- Matching strings that we should not have matched (there, then, other)
  - False positives (Type I)
- Not matching things that we should have matched (The)
  - False negatives (Type II)

Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment) positives	Actual Value (as confirmed by experiment) negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

https://alearningaday.blog/2016/09/14/confusion-matrix/

	Actual Value		
Predicted Value		Positives	Negatives
	Positives	the	there, then, other
	Negatives	The	There

Slide taken from Dr. Haithem Afli's Lecture on NLP

Confusion Matrix

		CONDITION determined by "Gold Standard"		
	TOTAL POPULATION	CONDITION POS	CONDITION NEG	PREVALENCE $\frac{\text{CONDITION POS}}{\text{TOTAL POPULATION}}$
TEST OUT-COME	TEST POS	True Pos TP	*Type I Error* False Pos FP	*Precision* Pos Predictive Value $PPV = \frac{TP}{TEST\ P}$
				False Discovery Rate $FDR = \frac{FP}{TEST\ P}$
	TEST NEG	*Type II Error* False Neg FN	True Neg TN	False Omission Rate $FOR = \frac{FN}{TEST\ N}$
				Neg Predictive Value $NPV = \frac{TN}{TEST\ N}$
	ACCURACY ACC $ACC = \frac{TP + TN}{TOT\ POP}$	*Sensitivity (SN), Recall* Total Pos Rate TPR $TPR = \frac{TP}{CONDITION\ POS}$	*Fall-Out* False Pos Rate FPR $FPR = \frac{FP}{CONDITION\ NEG}$	Pos Likelihood Ratio LR + $LR+ = \frac{TPR}{FPR}$
				Diagnostic Odds Ratio DOR $DOR = \frac{LR+}{LR-}$
		Miss Rate False Neg Rate FNR $FNR = \frac{FN}{CONDITION\ POS}$	*Specificity (SPC)* True Neg Rate TNR $TNR = \frac{TN}{CONDITION\ NEG}$	Neg Likelihood Ratio LR - $LR- = \frac{TNR}{FNR}$

https://www.unite.ai/what-is-a-confusion-matrix/

Confusion Matrix

- Set of words: {the, The, There, then, other, there}
- Ground Truth: {the, The}
- Regex(Set of words): {The, There}
- Identify Quadrant?
  - Q1
  - Q2
  - Q3
  - Q4
  - Type 1 error
  - Type 2 error

		Predicted Value	
Actual Value		Positives	Negatives
	Positives	Q1	Q2
	Negatives	Q3	Q4

Confusion Matrix

- Set of words: {the, The, There, then, other, there}
- Ground Truth: {the, The}
- Regex(Set of words): {The, There}
- Identify Quadrant?
  - Q1
  - Q2
  - Q3
  - Q4
  - Type 1 error
  - Type 2 error

		Predicted Value	
Actual Value		Positives	Negatives
	Positives	TP	FN/T2
	Negatives	FP/T1	TN

Confusion Matrix

- Set of words: {the, The, There, then, other, there}
- Ground Truth: {the, The}
- Regex(Set of words): {The, There}
- Identify Quadrant?
  - the
  - The
  - There
  - Then
  - other
  - there

		Predicted Value	
Actual Value		Positives	Negatives
	Positives	Q1	Q2
	Negatives	Q3	Q4

Confusion Matrix

- Set of words: {the, The, There, then, other, there}
- Ground Truth: {the, The}
- Regex(Set of words): {The, There}
- Identify Quadrant?
  - the
  - The
  - There
  - Then
  - other
  - there

	Actual Value		
Predicted Value		Positives	Negatives
	Positives	The	There
	Negatives	the	then, other, there

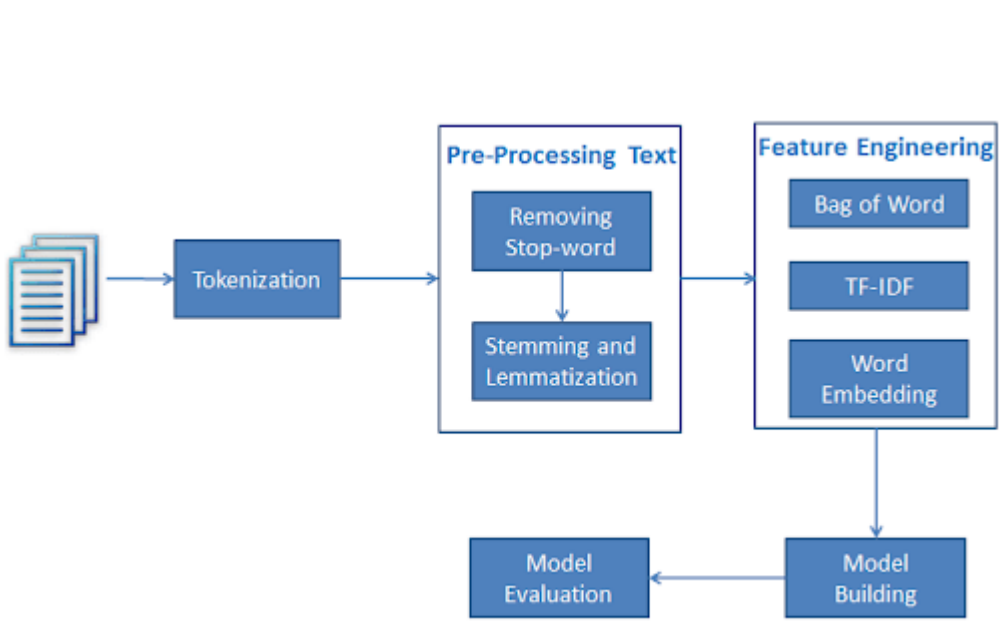
Resolution : Regular Expression

- In NLP we are always dealing with these kinds of errors.
- Reducing the error rate for an application often involves two antagonistic efforts:
  - Increasing accuracy or precision (minimizing false positives)
  - Increasing coverage or recall (minimizing false negatives).

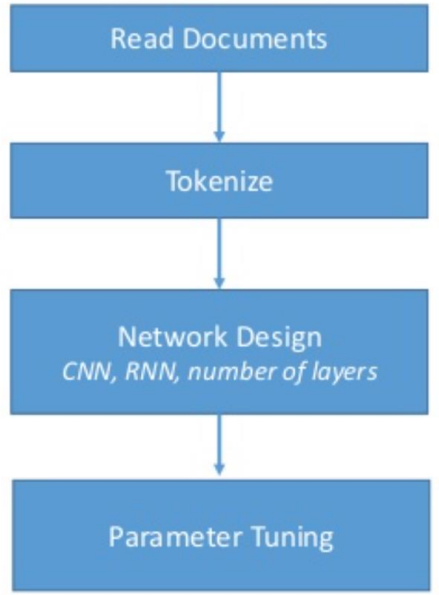
Slide taken from Dr. Haithem Afli's Lecture on NLP

Text classification

- Steps of Text Classification



Traditional Approach



Deep Learning Approach

https://www.slideshare.net/SomnathBanerjee17/classifying-text-with-cnn
https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk

Lab Exercise

- **Cygwin**

CORK INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA CHORCAÍ

http://www.cit.ie

Thank you

Praveen Joshi

05/10/2020