



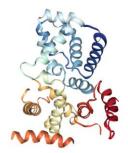


### **Example Problem**

Classify protein fold as either all alpha or all beta given the protein sequence.

MHHHHHHSSGRENLYFQGMTVREKTRLEKFRQLLSSQNTDLDELRKCS WPGVPREVRPITWRLLSGYLPANTERRKLTLQRKREEYFGFIEQYYDSR NEEHHQDTYRQIHIDIPRTNPLIPLFQQPLVQEIFERILFIWAIRHPASGYV QGINDLVTPFFVVFLSEYVEEDVENFDVTNLSQDMLRSIEADSFWCMSK LLDGIQDNYTFAQPGIQKKVKALEELVSRIDEQVHNHFRRYEVEYLQFAF RWMNNLLMRELPLRCTIRLWDTYQSEPEGFSHFHLYVCAAFLIKWRKEI LDEEDFQGLLMLLQNLPTIHWGNEEIGLLLAEAYRLKYMFADAPNHYRR





PDB ID: 3DZX - all alpha

GSSGSSGLPQVEAYSPSACSVRGGEELVLTGSNFLPDSKVVFIERGPD GKLQWEEEATVNRLQSNEVTLTLTVPEYSNKRVSRPVQVYFYVSNGRR KRSPTQSFRFLPVICKEE





PDB ID: 2YRP – all beta

# **PySpark Machine Learning**

- pyspark.ml
  - Uses Dataset
- pyspark.mllib (maintenance mode)
  - Uses RDD

#### **Feature Vector Creation**

(2-grams) 50-dimensional sequence words n-grams feature vector SALHWR...  $\Longrightarrow$  [S A, A L, L H, H...  $\Longrightarrow$  [-0.867...

Feature Engineering Pipeline: NGram Word2Vec

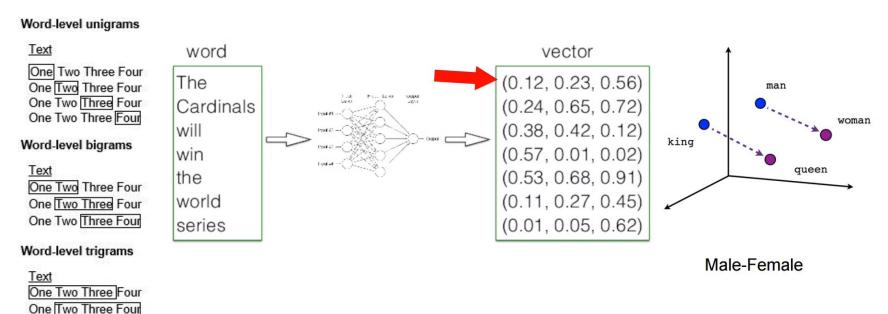
alpha  beta  coil foldType  words  ngram  features    0.7194245  0.0  0.28057554  alpha [S, A, L, H, W, R [S A, A L, L H, H [-0.8678886349164    0.4090909  0.0  0.59090906  alpha [S, H, L, K, S, K [S H, H L, L K, K [-0.2172588950821    0.015151516  0.5808081  0.4040404  beta [M, A, H, H, H, H [M A, A H, H H, H [-0.6485012823101    0.7916667  0.0  0.20833333  alpha [M, G, S, S, H, H [M G, G S, S S, S [-0.7672874573129    0.028846154  0.52884614  0.44230768  beta [M, E, K, A, T, K [M E, E K, K A, A [-0.7308713050851    0.6904762 0.011904762  0.29761904  alpha [P, T, I, H, D, H [P T, T I, I H, H [-0.7869252321949    0.8108108  0.0  0.1891892  alpha [G, A, M, E, P, E [G A, A M, M E, E [-0.8350848566740    0.022900764  0.39694658  0.5801527  beta [I, V, N, G, E, E [I V, V N, N G, G [-0.6840203473201	L														
0.7194245		alphal	betal	coillf	foldTypel			١	wordsl				ngraml	feature	s١
0.8021978      0.0  0.19 78022  alpha [M, T, P, D, V, L [M T, T P, P D, D [-0.7698143909806	+	0.7194245  0.4090909  0.015151516  0.7916667  0.028846154  0.6904762  0.8108108  0.022900764	0.0  0.0  0.5808081  0.0  0.52884614  0.011904762  0.0  0.39694658	0.28057554   0.59090906   0.4040404   0.20833333   0.44230768   0.29761904   0.1891892   0.5801527	alphal[S, Aalphal[S, Aalphal[M, Aalphal[M, Balphal[P, Talphal[G, Aalphal[I, Nalphal[I, N	A, L, H, L, A, H, G, S, E, K, T, I, A, M, V, N,	H, K, H, S, A, H, E, G,	W, S, H, T, D, P,	RI KI HI KI EI	[S A, [S H, [M A, [M G, [M E, [P T, [G A, [I V,	A L, H L, A H, G S, E K, T I, A M, V N,	L H, L K, H H, S S, K A, I H, M E, N G,	H  K  H  S  A  E  G	[-0.8678886349164 [-0.2172588950821 [-0.6485012823101 [-0.7672874573129 [-0.7308713050851 [-0.7869252321949 [-0.8350848566740 [-0.6840203473201	

https://github.com/sbl-sdsc/mmtfpyspark/blob/master/mmtfPyspark/ml/proteinSequenceEncoder.py



### N-gram, Word2Vec

- N-gram: Slice and group a string with a window size of N
- Word2Vec: Projecting a string to an ndimensional vector



https://spark.apache.org/docs/latest/mllib-feature-extraction.html#word2vec



### **Create Class Labels**

+	+	+		
	lphal	betal	coil	foldTypel
+	+	+		+
0.719	42451	0.01	0.28057554	alphal
1 0.409	09091	0.01	0.59090906	alpha
0.01515	1516  0.	5808081	0.4040404	betal
l 0.791	66671	0.01	0.20833333	alphal
1 0.02884	6154  0.5	28846141	0.44230768	betal
0.690	4762 0.01	19047621	0.29761904	alphal
0.810	8108	0.01	0.1891892	alphal
1 0.02290	0764  0.3	96946581	0.5801527	betal
1 0.802	1978	0.01	0.1978022	alphal

### The Dataset for Classification

We save this dataset as .parquet file:

```
data.write.mode("overwrite").format("parquet").save(filename);
```

class lahal

Class label +																		
I	alphal	betal	   coil foldType  +			wordsl					l ngram					l features		
<del> </del>	0.7194245	0.01 0.01	0.280575541	alpha [S, alpha [S,	Α,	L,	Н,	W,	R	I[S A	λ, Δ	\ L,	LH	, н	<b>[</b> -0	.867888	634916	54l
İ	0.015151516	0.5808081	0.4040404	betal[M,	Α,	Н,	Н,	Н,	H	[M A	λ, Δ	Н,	н н	, н	<b>[</b> -0	.648501	282310	)1l
l l	0.7916667 0.028846154	0.0l 0.528846141		alphal[M, betal[M,	•	•	•	•		_	•	•		•	_			
İ		0.011904762		alphal[P,		-				_	•				_			
I	0.8108108	0.01	0.1891892	alphal[G,	Α,	Μ,	Ε,	Ρ,	E	[G A	۸, ۵	М,	M E	, E	<b>[-0</b>	.835084	856674	۱۱
۱	0.0229007641	0.396946581	0.5801527	betal[I,	۷,	N,	G,	Ε,	E	I[I V	/, V	/ N,	N G	, G	I [-Ø	.684020	347320	)1l
П	0.80219781	0.0	0.19 <sub>78022</sub> 1	alpha∣[M,	Τ,	Ρ,	D,	٧,	L	ІГМ Т	Т, Т	P,	P D	. D	I Γ-0	769814	390980	<b>3</b> 61

fasturas

### **Demos**

Create a dataset

Fit several classification models with pyspark.ml

Fit several classification models with scikit-learn

#### **Problem 1**

- Change the code to a 3-state classification problem:
  - alpha, beta, alpha+beta
- Rerun the classification with Decision Tree Classifier

### Resources

- Pyspark.ml API
  - http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html
- Extracting, transforming and selecting features
  - http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#modulepyspark.ml.feature
  - N-gram
    - http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#pyspark.ml.feature.NGram
  - Word2Vec
    - http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html#pyspark.ml.feature.
       Word2Vec
- Classification and regression
  - https://spark.apache.org/docs/2.2.0/ml-classification-regression.html
- Parquet files (columnar format)
  - https://spark.apache.org/docs/latest/sql-programming-guide.html#parquet-files





# **Funding**

This workshop was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

