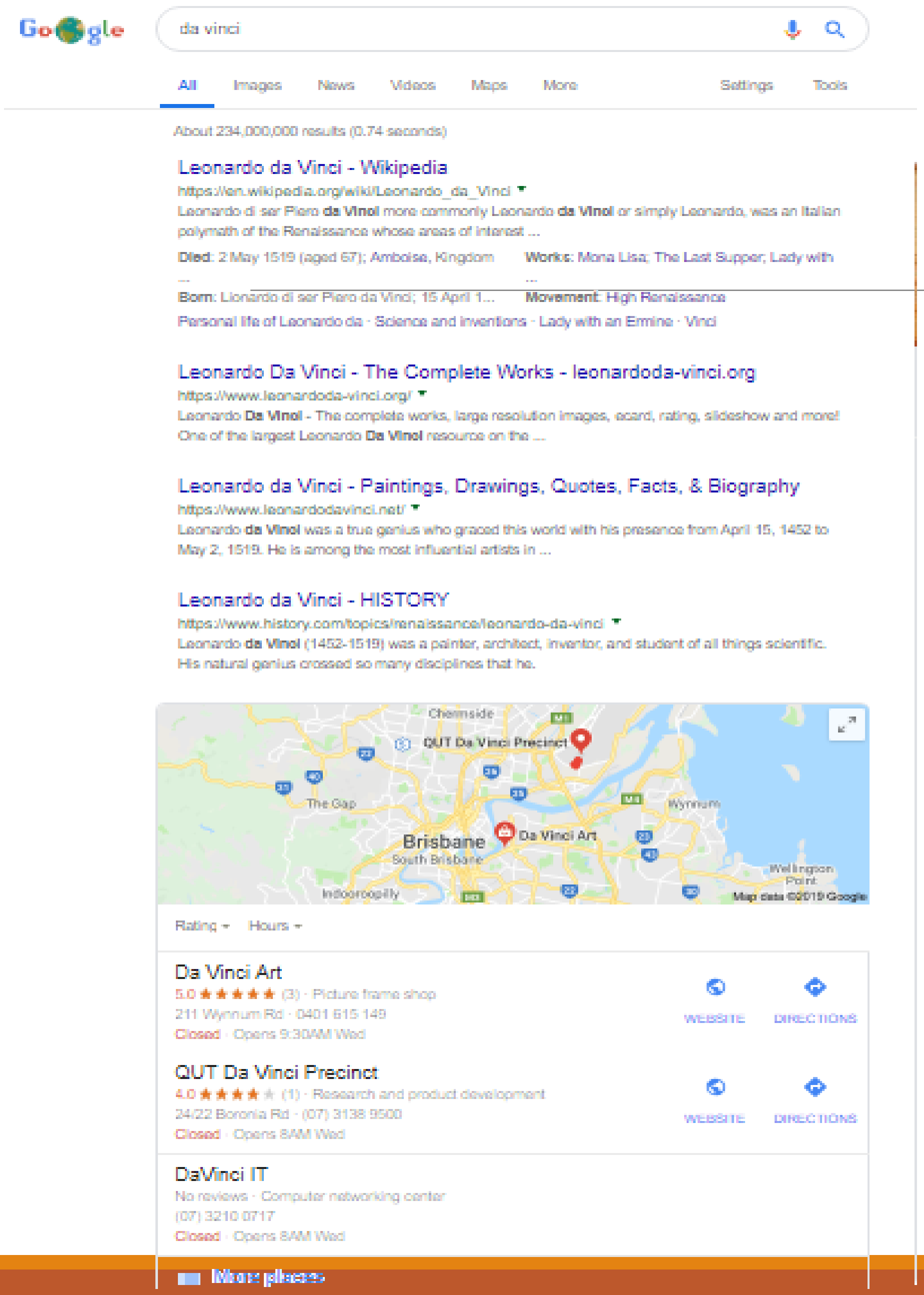
IFN702 Research Project

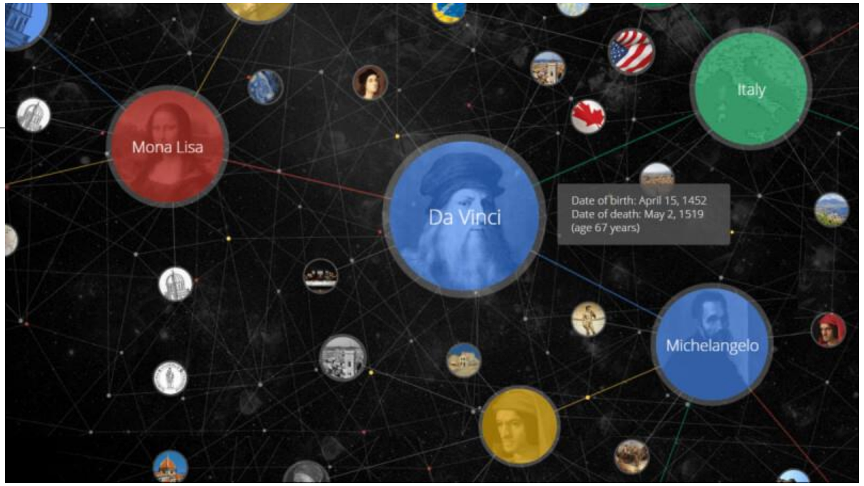
Machine learning by tensor base model approach to infer new relations in the YAGO knowledge graph

Mary Rose Legaspi (n10086820)

Mr Monir Moniruzzaman (Supervisor)

Google Knowledge Graph



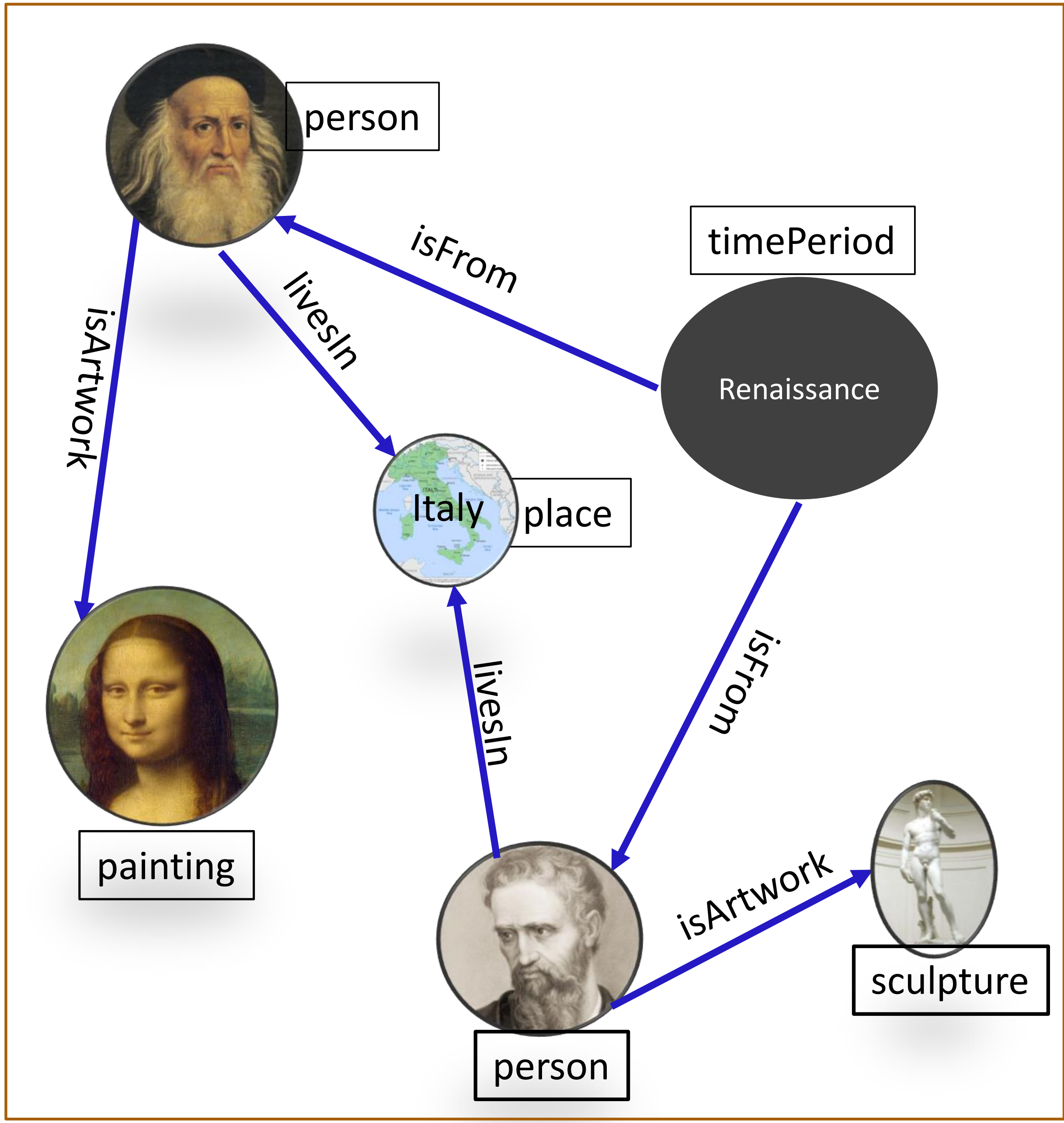


Knowledge panel

Vital information about people, places, things etc.

Knowledge Graph

- Knowledge in a form of graph
- Entities, attributes, and relationships
- Entities are the nodes
- Nodes are labelled with attributes
- Relationship between entities are link through types edges



Who uses knowledge graph

- Google knowledge graph
- Facebook
- YouTube
- Amazon, etc.

Knowledge bases

- Freebase
- DbPedia
- YAGO

Motivation

- KG stores millions of entities and relations (edges and vertices) and contains billions of real world information.
- Despite, the extremely large size, the information is seemingly incomplete. There are missing relations.
- Information is growing exponentially at a rapid rate.

Purpose

- To extend the knowledge in the YAGO knowledge graph using its own knowledge base.

Scope

- Machine learning approach on knowledge graph to extract new knowledge to YAGO knowledge base.
- Apply tensor-based model on YAGO to infer new relations.

Deliverables

- Develop a tensor-based model using Matlab and Python programming language for inferring new relations between entities.
- Develop evaluation model to establish correctness of new relations to entities.
- List of newly discovered links using the model.
- Final written report.

Significance

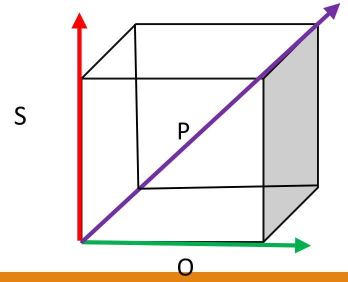
- Comprehensive information about the searched entity.
  - a detailed summary
- Improve user experience-sufficiently fulfill users' information needs.
  - Sufficient web search (time and quality).
- Fueling discovery – ability to foster new knowledge.

Literature

- Although many current large knowledge graphs have been published with millions of entities, missing or incorrect information are still present (Zhao, et al. 2017). As a result, several approaches have been made to handle the issue of knowledge graph completion.

- Latent feature models have been picking up pace in solving the problem of inferring missing relation in the knowledge graph. Neural Embedding methods and Factorisation are the two existing methods of latent feature models (Padia, Kalpakis, Ferraro, & Finin, 2019) that have shown incredible results.

- Factorisation methods use tensor, a three dimensional array, to represent triples in the knowledge graph and determine the probability of relations existing between entities (subject, object) though tensor factorisation (TF). Relationships between entities (nodes) can be derived from correlations of their latent factors is the main insight behind *Latent Feature Models.*

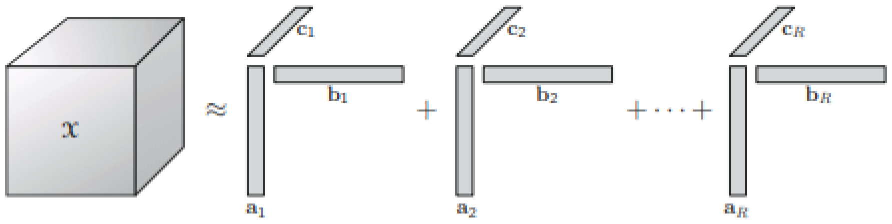


Literature cont.

- Popular tensor decomposition methods are CP decomposition and Tucker decomposition. Kolda and Bader (2009), stated that CANDECOMP/PARAFAC (CP) decomposition get the sum of rank-one tensors while Tucker decomposition is a higher order principal component analysis (PCA).

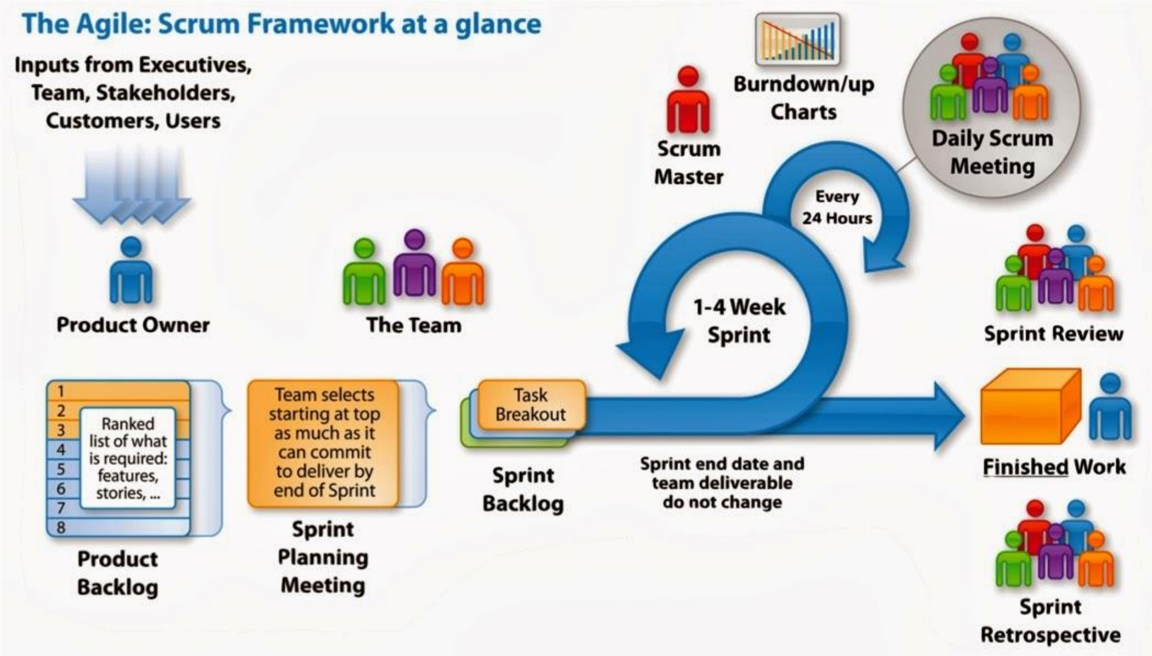
- The CP decomposition method is popularised by Kolda and Bader 2009 but work on similar topic has been originated much earlier by Hitchcock (1927) and Cattel (1944). In CP decomposition, it factorises the tensor into the sum of the outer components of the tensor.

- CP decomposition is a well known tensor decomposition technique use for ranking triples in knowledge graphs. Traditionally, CP is computed using ALS (alternating least square algorithm (Kolda & Bader 2009). Minimisation problems is attained by fixing all components except for one factor in the matrix sequentially and repeated until the criteria for stopping is met.



$$\mathcal{X} \approx [\![\mathbf{A},\mathbf{B},\mathbf{C}]\!] \equiv \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r .$$

Project Management: Agile Scrum technique



[Image courtesy of glurgeek.com]()

- **Adaptability**
- **Visibility**
- **Efficiency**

Tools and Technology Use

- Matlab
- TensorToolbox
- Python
- Jupyter Notebook
- Numpy
- Sckitlearn

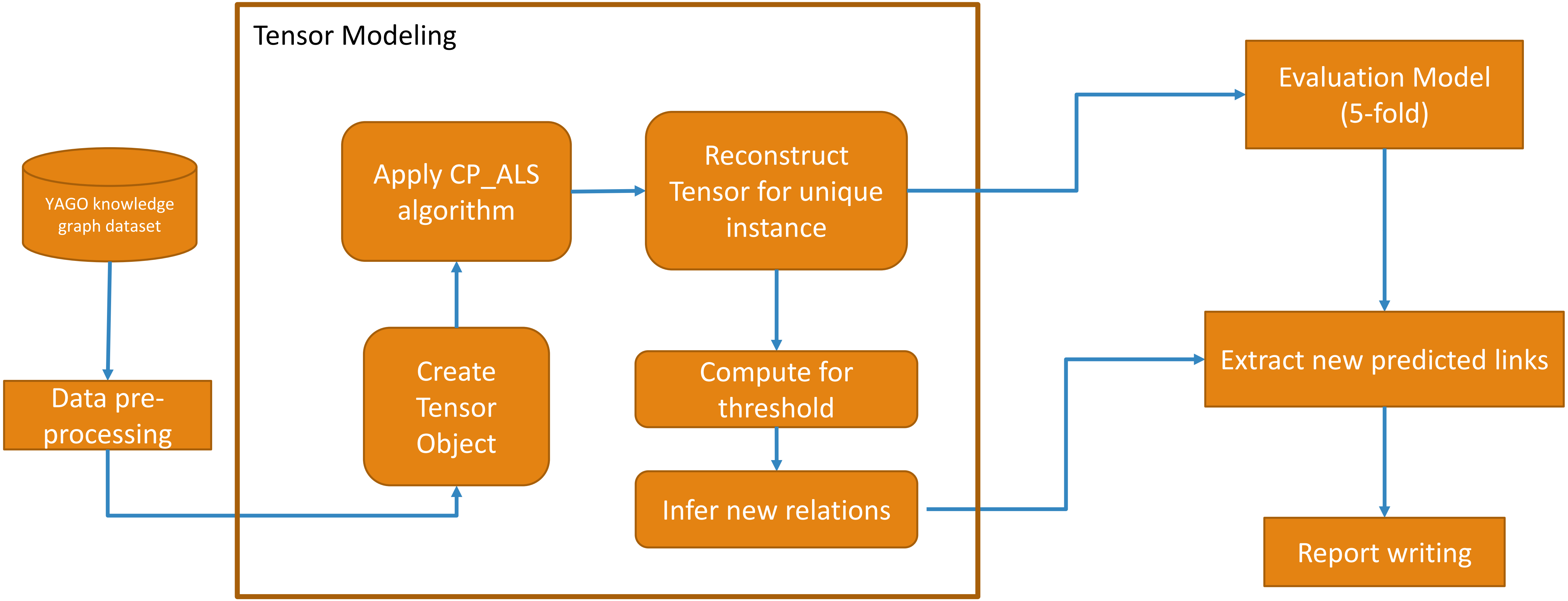
Dataset

```
@base <http://yago-knowledge.org/resource/> .
@prefix dbp: <http://dbpedia.org/ontology/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<yagoTheme_yagoDateFacts>       <hasGloss>      "This file is part of the ontology YAGO3. It is li
Institute for Informatics/Germany. See http://yago-knowledge.org for all details. This file was ge
<1st_Light_Car_Patrol_(Australia)>      <wasDestroyedOnDate>    "1919-##-##"^^xsd:date .
<Save_Chase_Farm>       <wasDestroyedOnDate>    "2010-##-##"^^xsd:date .
<Wanna-Ben>     <wasDestroyedOnDate>    "2012-03-23"^^xsd:date .
<Southern_Railway_(U.S.)>       <wasDestroyedOnDate>    "1990-##-##"^^xsd:date .
<The_Greatest_American_Hero>    <wasDestroyedOnDate>    "1983-02-03"^^xsd:date .
<Jagdstaffel_44>        <wasDestroyedOnDate>    "1918-##-##"^^xsd:date .
<5-4-3-2-Run>   <wasDestroyedOnDate>    "1990-##-##"^^xsd:date .
<Landherrnamt>  <wasDestroyedOnDate>    "1945-##-##"^^xsd:date .
<D_Records>     <wasDestroyedOnDate>    "1965-##-##"^^xsd:date .
<No._353_Squadron_RAF>  <wasDestroyedOnDate>    "1946-##-##"^^xsd:date .
<es/La_UniĆ³n_(La_Libertad,_PerĆ̦)>      <wasDestroyedOnDate>    "1973-##-##"^^xsd:date .
<Kuch_Toh_Log_Kahenge>  <wasDestroyedOnDate>    "2013-03-28"^^xsd:date .
<Care_Kori_Na>  <wasDestroyedOnDate>    "2013-03-09"^^xsd:date .
<Bezirk_Dresden>        <wasDestroyedOnDate>    "1990-##-##"^^xsd:date .
<de/Wiener_Katzen-Musik>        <wasDestroyedOnDate>    "1848-##-##"^^xsd:date .
<Come_with_Me_(TV_series)>      <wasDestroyedOnDate>    "2016-06-25"^^xsd:date .
<Alice_(season_3)>      <wasDestroyedOnDate>    "1979-04-01"^^xsd:date .
<The_Stillroven>        <wasDestroyedOnDate>    "1969-##-##"^^xsd:date .
<Market_Weighton_railway_station>       <wasDestroyedOnDate>    "1965-##-##"^^xsd:date .
<Bute_ministry> <wasDestroyedOnDate>    "1763-##-##"^^xsd:date .
<The_Hive_(website)>    <wasDestroyedOnDate>    "2004-##-##"^^xsd:date .
<Gwalior_State> <wasDestroyedOnDate>    "1948-##-##"^^xsd:date .
<1st/9th_Battalion,_Ulster_Defence_Regiment>    <wasDestroyedOnDate>    "1992-##-##"^^xsd:date .
<RFA_Sir_Percivale_(L3036)>     <wasDestroyedOnDate>    "2004-08-17"^^xsd:date .
<Round_Oak_railway_station>     <wasDestroyedOnDate>    "1962-##-##"^^xsd:date .
<Kuchh_Is_Tara> <wasDestroyedOnDate>    "2008-08-28"^^xsd:date .
<F.C._Ararat_Tehran>    <wasDestroyedOnDate>    "2009-##-##"^^xsd:date .
<RAF_Stoney_Cross>      <wasDestroyedOnDate>    "1946-##-##"^^xsd:date .
<The_Ellen_Burstyn_Show>        <wasDestroyedOnDate>    "1987-09-05"^^xsd:date .
<Norsk_Forurensningskontroll>   <wasDestroyedOnDate>    "1986-##-##"^^xsd:date .
```

- Yago KG subset (yagoFacts)
- Over 12 million triples
- 4.3 million entities
- 37 relations
- 4.3m x 37 x 4.3m triples

Over all Project methodology



Findings

Data pre-processing

Subject(S) Predicate (P) Object(O)

= Triple

	Unnamed: 0	<yagoTheme_yagoFacts>	<hasGloss>	This file is part of the ontology YAGO3. It is licensed under a Creative-Commons Attribution License by the YAGO team at the Max Planck Institute for Informatics/Germany. See http://yago-knowledge.org for all details. This file was generated on 2017-06-19 T 08:01:29.0836. All facts of YAGO that hold between instances CORE	Unnamed: 4
0	<id_rB6isMnplh_H?S_LT?Qo1Fzcl>	<Jesús_Rivera_Sánchez>	<isLeaderOf>	<Pueblo_of_Naranjito>	NaN
1	<id_BOK!FvTDPu_H?S_1NVSTKkFbS>	<Elizabeth_II>	<isLeaderOf>	<Royal_Numismatic_Society>	NaN
2	<id_Uy5EwU3nX1_H?S_otuJrkvKs1>	<Richard_Stallman>	<isLeaderOf>	<Free_Software_Foundation>	NaN
3	<id_A?rlHtKpyX_H?S_Ap3TBzfE6b>	<Keith_Peterson>	<isLeaderOf>	<Cambridge_Bay>	NaN
4	<id_vzhzgmCR5Y_H?S_9xW07qYiaH>	<William_H._Seward_Jr.>	<isLeaderOf>	<9th_New_York_Heavy_Artillery_Regiment>	NaN
5	<id_GqUh9jFAN?_H?S_A39fu5FWu4>	<Andranik>	<isLeaderOf>	<Armenian_fedayi>	NaN
6	<id_s60Psk1DHb_H?S_OACCn8W8Kv>	<Ramasamy_Palanisamy>	<isLeaderOf>	<Democratic_Action_Party_(Malaysia)>	NaN
7	<id_pII60Mnz8o_H?S_8mvRWxKXDG>	<Matt_Bevin>	<isLeaderOf>	<Kentucky_Air_National_Guard>	NaN
8	<id_losV58WRWE_H?S_KxmZk2LtbV>	<Leonard_Leo>	<isLeaderOf>	<Federalist_Society>	NaN

After Pre-processing

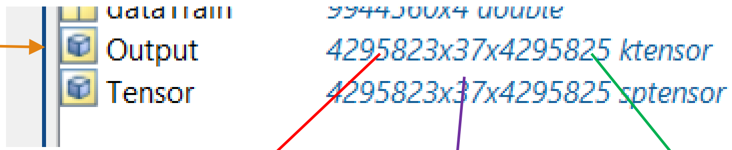
```
In [260]:  1  df1.head(20)
Out[260]:
```

	1347917	27	2223283
0	815813	27	2371560
1	2307361	27	948800
2	1510384	27	458907
3	2985292	27	44247
4	173060	27	227689
5	2260643	27	697471
6	1806594	27	1519146
7	1634987	27	887150
8	1305806	27	333693
9	163469	27	1531298
10	2829218	27	1917037
11	3036890	27	2750532
12	1761963	27	160566
13	2076027	27	146234
14	1764355	27	1066605
15	894376	27	228448

Findings

CP Decomposition
CP_ALS

```
Subs= eData(:,1:3);
Val = eData(:,4);
Tensor = sptensor(Subs, Val);
r = 4;
Output = cp_als(Tensor, r);
```



Output.u{1} Output.u{2} Output.u{3}

Findings

CP_ALS

In [260]:
```
1 df1.head(20)
```

Out[260]:

	1347917	27	2223283
0	815813	27	2371560
1	2307361	27	948800
2	1510384	27	458907
3	2985292	27	44247
4	173060	27	227689
5	2260643	27	697471
6	1806594	27	1519146
7	1634987	27	887150
8	1305806	27	333693
9	163469	27	1531298
10	2829218	27	1917037
11	3036890	27	2750532
12	1761963	27	160566
13	2076027	27	146234
14	1764355	27	1066605
15	894376	27	228448

Output.u{1}

	-5.963618433464708837e-271	4.145256649836981368e-271	1.513342989837988009e-268	-1.023224598487109372e-267
0	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
1	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
2	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
3	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
4	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
5	7.097201e-04	8.615601e-05	1.223030e-06	8.963834e-06
6	7.097214e-04	8.615515e-05	-7.616006e-07	1.136198e-05
7	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
8	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
9	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
10	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
11	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
12	-1.796468e-15	1.188610e-15	9.458197e-13	-3.318294e-12
13	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
14	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
15	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
16	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
17	-7.996197e-12	5.572305e-12	-2.930259e-09	-1.354924e-08
18	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
19	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
20	-1.323445e-10	7.394752e-11	5.966212e-07	-3.042222e-07
21	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
22	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
23	-1.107075e-10	7.500643e-11	-8.265036e-08	-1.970127e-07
24	-7.126533e-12	4.721871e-12	6.427005e-09	-1.314468e-08
25	-1.466790e-13	1.026557e-13	-6.594561e-11	-2.466106e-10
26	-6.865732e-12	4.533929e-12	6.542129e-09	-1.272994e-08
27	-6.865732e-12	4.533929e-12	6.542129e-09	-1.272994e-08
28	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
29	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00
...	...	...	...	...
4295792	-9.639489e-13	5.594199e-13	3.793729e-09	-2.124659e-09
4295793	-3.518823e-18	1.957113e-18	1.627072e-14	-8.128802e-15
4295794	0.000000e+00	0.000000e+00	0.000000e+00	-0.000000e+00

Output.u{1} represents the latent factors for the 4295823 subject entity.

Findings

CP_ALS cont.

In [260]:

```
df1.head(20)
```

Out[260]:

	1347917	27	2223283
0	815813	27	2371560
1	2307361	27	948800
2	1510384	27	458907
3	2985292	27	44247
4	173060	27	227689
5	2260643	27	697471
6	1806594	27	1519146
7	1634987	27	887150
8	1305806	27	333693
9	163469	27	1531298
10	2829218	27	1917037
11	3036890	27	2750532
12	1761963	27	160566
13	2076027	27	146234
14	1764355	27	1066605
15	894376	27	228448

.
.
.

Output.u{2}

	-4.9854631857817349047e-15	-3.326300082887661847e-16	-6.168347611941683076e-14	1.6527752115509370852e-12
0	2.527797e-09	2.916461e-09	7.233712e-08	1.401754e-06
1	9.606957e-12	1.322033e-12	1.655859e-09	-2.082152e-09
2	-5.874671e-06	1.790434e-07	1.212826e-03	1.509670e-03
3	-1.059968e-15	2.368471e-16	-1.796425e-14	4.243317e-13
4	-2.423834e-16	3.915455e-17	-6.131820e-15	1.096626e-13
5	3.419283e-68	3.965536e-69	-6.344337e-67	-6.453097e-66
6	-1.450093e-16	2.717947e-17	-4.504678e-15	7.879463e-14
7	1.165883e-10	1.498443e-11	1.074752e-08	-2.389762e-08
8	2.393937e-20	1.290671e-20	-4.917776e-18	5.460046e-17
9	3.438724e-15	5.350477e-16	1.138568e-12	-8.253590e-13
10	-2.940404e-18	-1.066634e-19	-5.383544e-17	1.109369e-15
11	3.601176e-19	4.319233e-20	7.438777e-18	-6.994849e-17
12	9.999908e-01	1.000000e+00	-7.722341e-04	-1.351628e-02
13	-6.180946e-59	1.909555e-59	-7.708070e-57	3.083098e-56
14	5.799605e-14	6.856086e-15	1.434123e-13	-1.112071e-11
15	3.488117e-22	4.016088e-23	-9.124171e-21	-6.544484e-20
16	8.839507e-118	1.204408e-118	5.897143e-115	-3.417450e-115
17	-4.466608e-16	5.880010e-17	3.587282e-16	1.646864e-13
18	2.461158e-72	2.864241e-73	-3.691226e-71	-4.657705e-70
19	-8.380071e-18	1.269050e-18	-6.596878e-17	3.432726e-15
20	-6.511235e-14	4.776690e-15	6.512116e-12	2.020180e-11
21	-4.295478e-03	1.583667e-05	9.999056e-01	9.998265e-01
22	3.494131e-15	4.397623e-16	2.426338e-13	-7.044006e-13
23	-1.916793e-10	9.310105e-11	6.803953e-09	1.986293e-08
24	-2.322328e-11	-9.870691e-12	-2.563290e-11	1.454592e-09
25	-9.767602e-08	-2.675472e-09	8.269642e-06	2.152646e-05
26	2.448664e-07	4.828108e-08	1.708965e-04	-7.194903e-05
27	-3.724869e-18	3.738133e-19	-6.387567e-17	1.615170e-15
28	-1.494054e-05	6.760847e-07	6.366494e-03	3.226943e-03
29	-1.847173e-05	1.006428e-06	6.033435e-03	4.618779e-03
30	5.800627e-12	8.853392e-13	1.758465e-09	-1.366022e-09
31	-1.305528e-17	-2.802020e-18	3.819037e-17	1.873306e-15
32	-2.106344e-16	-4.535431e-18	6.141673e-16	5.546814e-14
33	-4.392447e-05	2.241502e-06	1.047395e-02	1.141268e-02
34	-5.162209e-16	8.330975e-18	6.377772e-15	1.534428e-13
35	-5.864778e-16	1.189483e-16	-9.818955e-15	2.470579e-13

Output.u{2} represents the latent factors for the 37 predicate(triple relations)

Findings

CP_ALS cont.

In [260]:
```
df1.head(20)
```

Out[260]:

	1347917	27	2223283
0	815813	27	2371560
1	2307361	27	948800
2	1510384	27	458907
3	2985292	27	44247
4	173060	27	227689
5	2260643	27	697471
6	1806594	27	1519146
7	1634987	27	887150
8	1305806	27	333693
9	163469	27	1531298
10	2829218	27	1917037
11	3036890	27	2750532
12	1761963	27	160566
13	2076027	27	146234
14	1764355	27	1066605
15	894376	27	228448

Object entity (O)
latent factor

Output.u{3}

	0	0.1	0.2	0.3
0	-1.154863e-278	-1.782555e-279	4.021566e-277	-3.296939e-276
1	7.660132e-15	-5.310940e-15	-5.315445e-14	2.086942e-12
2	1.591491e-15	1.147856e-15	-9.054625e-16	1.559539e-14
3	9.755955e-22	3.207256e-23	-6.993181e-20	3.799787e-19
4	-2.626169e-24	-4.015733e-25	1.129750e-22	-7.537698e-22
5	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
6	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
7	7.660149e-15	-5.310939e-15	-5.315471e-14	2.086948e-12
8	1.189279e-31	1.835728e-32	-4.138374e-30	3.395131e-29
9	4.132364e-20	3.003149e-21	-2.322134e-18	1.551912e-17
10	1.550250e-15	1.141316e-15	-4.694388e-16	4.009860e-15
11	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
12	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
13	-1.266000e-278	-1.954096e-279	4.408575e-277	-3.614215e-276
14	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
15	1.720468e-14	-3.840672e-15	-4.033043e-13	4.815106e-12
16	9.076115e-15	2.339489e-15	-5.426115e-14	2.113382e-12
17	1.672955e-15	1.160353e-15	-4.164723e-15	3.893321e-14
18	3.097121e-15	1.380405e-15	-5.257278e-14	4.452857e-13
19	9.572508e-15	2.415941e-15	-7.246235e-14	2.255270e-12
20	3.056432e-14	8.434140e-15	-2.327385e-13	6.855323e-12
21	8.982249e-15	2.324618e-15	-5.315784e-14	2.086994e-12
22	1.574619e-15	1.145121e-15	-1.064390e-15	1.091897e-14
23	1.913356e-15	1.197450e-15	-1.260019e-14	1.075771e-13
24	1.390223e-22	2.503805e-23	1.545274e-20	3.588868e-20
25	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00

Output.u{1} represents the latent factors for the 4295823 object entity.

Findings

CP_ALS Output

```
In [260]:  1  df1.head(20)
Out[260]:
```

	1347917	27	2223283
0	815813	27	2371560
1	2307361	27	948800
2	1510384	27	458907
3	2985292	27	44247
4	173060	27	227689
5	2260643	27	697471
6	1806594	27	1519146
7	1634987	27	887150
8	1305806	27	333693
9	163469	27	1531298
10	2829218	27	1917037
11	3036890	27	2750532
12	1761963	27	160566
13	2076027	27	146234
14	1764355	27	1066605
15	894376	27	228448

```
>> Output(815813, 27, 2371560)

ans =

   5.4758e-18
```

The output value for the triple<818513> <27> <2371560> is 5.4758e-18.
The output value is the sum of the latent factors (s, p, o) .

Findings

“Ellen_Greene” actedIn Inference

Missing Object Inference

<subject, e> <predicate> <???>

403577	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(1919_film)>"	818027	1	2594595	1.0192e-15	0
403578	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(1931_film)>"	818027	1	2594596	1.0192e-15	0
403579	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(1991_film)>"	818027	1	2594597	4.1007e-15	1
403580	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(Diana_Krall_album)>"	818027	1	2594598	1.8218e-15	0
403581	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(Steve_Laury_album)>"	818027	1	2594600	1.9717e-18	0
403582	<Ellen_Greene>	<actedIn>	"<Stepping_Out_(play)>"	818027	1	2594602	4.1007e-15	0
403583	<Ellen_Greene>	<actedIn>	"<Stepping_Selection>"	818027	1	2594604	1.9329e-33	0
403584	<Ellen_Greene>	<actedIn>	"<Stepping_Sisters>"	818027	1	2594605	4.0902e-15	0
403585	<Ellen_Greene>	<actedIn>	"<Stepping_Stone_(Duffy_song)>"	818027	1	2594607	5.0998e-15	0
403586	<Ellen_Greene>	<actedIn>	"<Stepping_Stone_(Jimi_Hendrix_song)>"	818027	1	2594608	1.0192e-15	0
403587	<Ellen_Greene>	<actedIn>	"<Stepping_Stone_(Lari_White_song)>"	818027	1	2594609	2.0383e-15	0
403588	<Ellen_Greene>	<actedIn>	"<Stepping_Stone_(album)>"	818027	1	2594610	1.0191e-15	0
403589	<Ellen_Greene>	<actedIn>	"<Stepping_Stones_(album)>"	818027	1	2594615	6.7051e-16	0
403590	<Ellen_Greene>	<actedIn>	"<Stepping_Stones_(musical)>"	818027	1	2594619	4.5934e-22	0
403591	<Ellen_Greene>	<actedIn>	"<Stepping_Toes>"	818027	1	2594622	3.2082e-15	0
403592	<Ellen_Greene>	<actedIn>	"<Stepping_into_Tomorrow>"	818027	1	2594623	1.0192e-15	0
403593	<Ellen_Greene>	<actedIn>	"<Stepping_on_Angels..._Before_Dawn>"	818027	1	2594624	-1.9333e-277	0
403594	<Ellen_Greene>	<actedIn>	"<Stepping_on_the_Crowtche_owf_Your_A...	818027	1	2594625	1.9717e-18	0
403595	<Ellen_Greene>	<actedIn>	"<Stepps>"	818027	1	2594628	-6.2389e-163	0
403596	<Ellen_Greene>	<actedIn>	"<Steps_(book)>"	818027	1	2594633	1.1022e-15	0
403597	<Ellen_Greene>	<actedIn>	"<Steps_Ahead_(album)>"	818027	1	2594636	1.0191e-15	0
403598	<Ellen_Greene>	<actedIn>	"<Steps_discography>"	818027	1	2594637	1.1738e-30	0
403599	<Ellen_Greene>	<actedIn>	"<Steps_in_Time>"	818027	1	2594638	6.5497e-32	0

Positive triple

Recommended missing triple

Here in row 403579, the triple <Ellen_Greene> <actedIn><Stepping_Out_(1991_film)> is a positive triple with latent factor 4.1007e-15.
We can infer that row number 403582 <Ellen_Greene> <actedIn><Stepping_Out_(play)> could be a possible candidate as recommended missing object entity with latent factor value similar of that of the positive triple.

Findings

Missing Object Inference

“Tom_Shadyac” directed Inference

<subject, e> <predicate> <???>

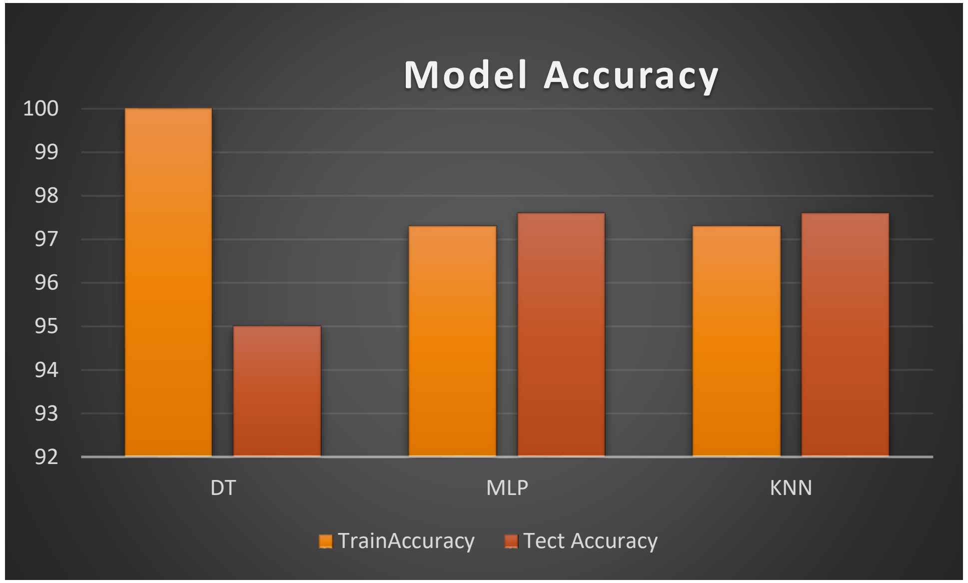
383610	<Tom_Shadyac>	<directed>	"<Liar,_Liar,_Vampire>"	2792153	5	1648580	3.9342e-22	0
383611	<Tom_Shadyac>	<directed>	"<Liar,_Liar_(1993_film)>"	2792153	5	1648581	6.6892e-23	0
383612	<Tom_Shadyac>	<directed>	"<Liar,_Liar_(The_Castaways_song)>"	2792153	5	1648582	3.2515e-16	0
383613	<Tom_Shadyac>	<directed>	"<Liar/Dead_Is_the_New_Alive>"	2792153	5	1648583	6.6289e-16	0
383614	<Tom_Shadyac>	<directed>	"<Liar_&_a_Thief>"	2792153	5	1648584	2.6309e-16	0
383615	<Tom_Shadyac>	<directed>	"<Liar_(Eskimo_Joe_song)>"	2792153	5	1648585	8.6504e-16	0
383616	<Tom_Shadyac>	<directed>	"<Liar_(Fake_Shark_â€"_Real_Zombie!_al...	2792153	5	1648586	-2.2381e-287	0
383617	<Tom_Shadyac>	<directed>	"<Liar_(Harisu_album)>"	2792153	5	1648587	1.0521e-18	0
383618	<Tom_Shadyac>	<directed>	"<Liar_(Profyle_song)>"	2792153	5	1648588	5.2567e-16	0
383619	<Tom_Shadyac>	<directed>	"<Liar_(Queen_song)>"	2792153	5	1648589	1.0904e-15	0
383620	<Tom_Shadyac>	<directed>	"<Liar_(Rollins_Band_song)>"	2792153	5	1648590	-9.2042e-274	0
383621	<Tom_Shadyac>	<directed>	"<Liar_(Russ_Ballard_song)>"	2792153	5	1648591	9.0908e-16	0
383622	<Tom_Shadyac>	<directed>	"<Liar_(The_Jesus_Lizard_album)>"	2792153	5	1648592	2.7968e-32	0
383623	<Tom_Shadyac>	<directed>	"<Liar_(novel)>"	2792153	5	1648593	3.9371e-17	0
383624	<Tom_Shadyac>	<directed>	"<Liar_Game>"	2792153	5	1648594	5.0992e-19	0
383625	<Tom_Shadyac>	<directed>	"<Liar_Liar>"	2792153	5	1648597	2.6446e-16	1
383626	<Tom_Shadyac>	<directed>	"<Liar_Liar_(Cris_Cab_song)>"	2792153	5	1648598	7.9268e-16	0
383627	<Tom_Shadyac>	<directed>	"<Liar_Wanted>"	2792153	5	1648599	7.7336e-18	0
383628	<Tom_Shadyac>	<directed>	"<Liarbird>"	2792153	5	1648600	5.0157e-17	0
383629	<Tom_Shadyac>	<directed>	"<Liard_Island>"	2792153	5	1648601	-2.3904e-30	0
383630	<Tom_Shadyac>	<directed>	"<Liari>"	2792153	5	1648606	3.0460e-14	0
383631	<Tom_Shadyac>	<directed>	"<Liars_(Liars_album)>"	2792153	5	1648608	-2.8507e-276	0
383632	<Tom_Shadyac>	<directed>	"<Liars_(Todd_Rundgren_album)>"	2792153	5	1648609	2.9142e-16	0

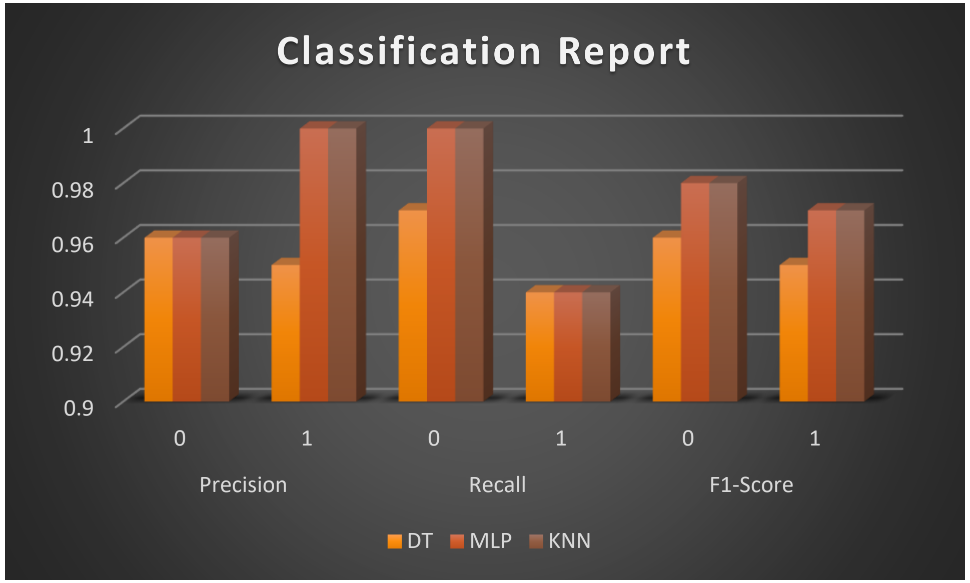
Recommended missing triple

Positive triple

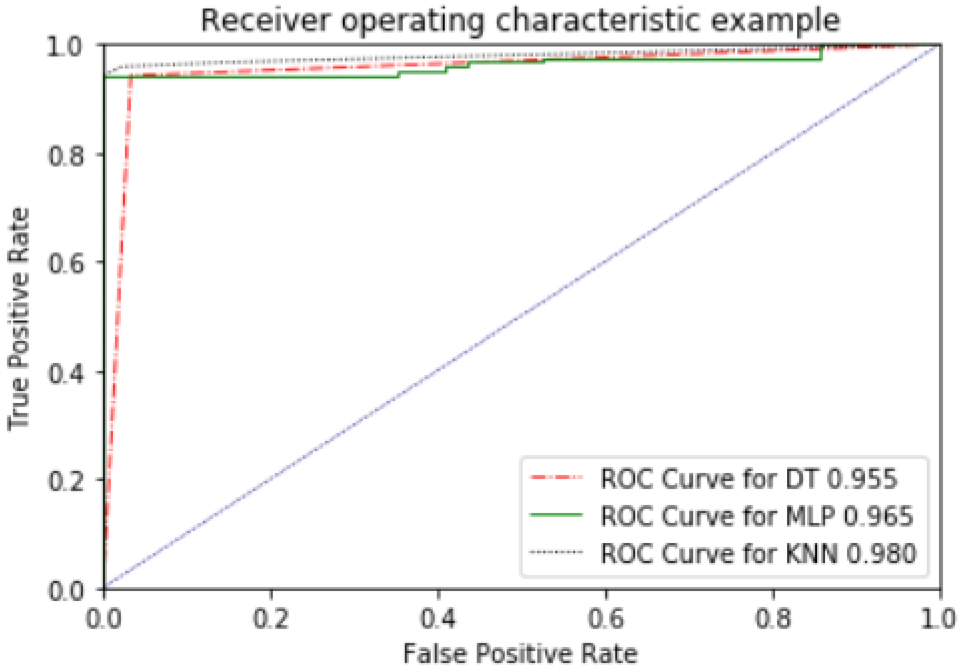
Here in the triple <Tom_Shaydac><directed><Liar_Liar> is a positive triple with latent 2.6446e-16. We can infer that row 353614 could be a possible candidate as recommended missing object entity with latent factor value of 2.6309e-16 which is closer to the positive original triple. Thus, inference can be suggested.

Evaluation (Post processing)

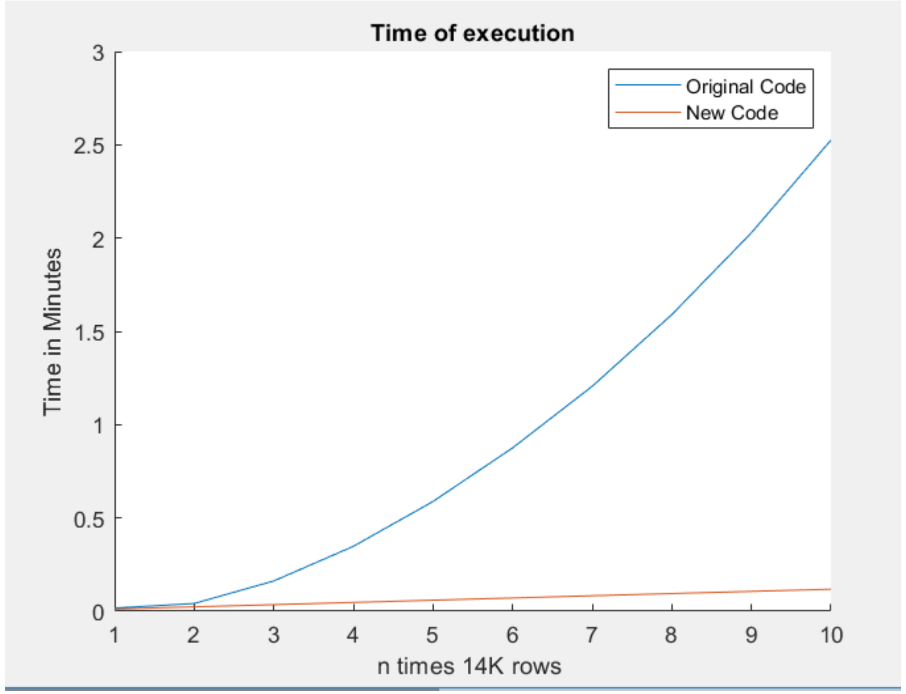




Evaluation


Receiver operating characteristic example
True Positive Rate
False Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve for DT 0.955
ROC Curve for MLP 0.965
ROC Curve for KNN 0.980

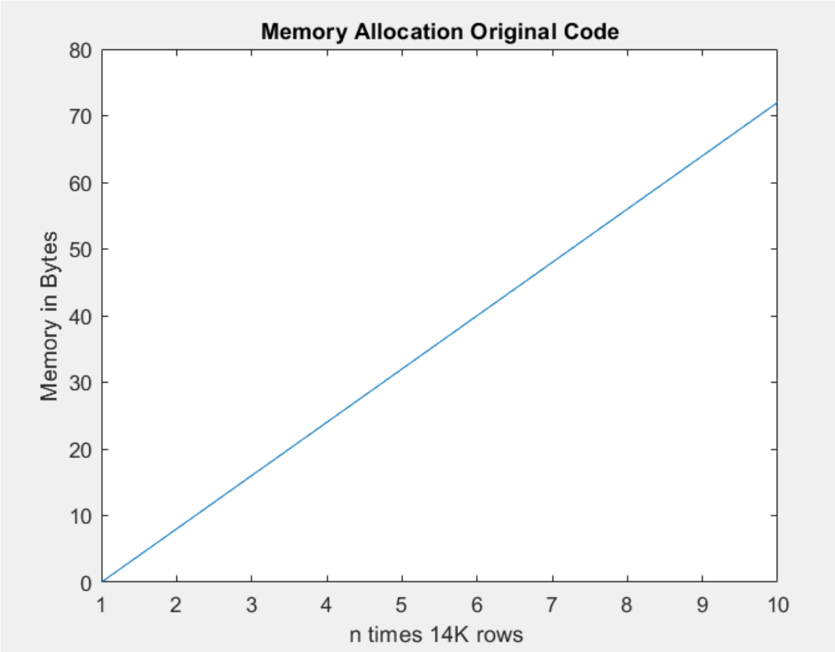
Run time and memory

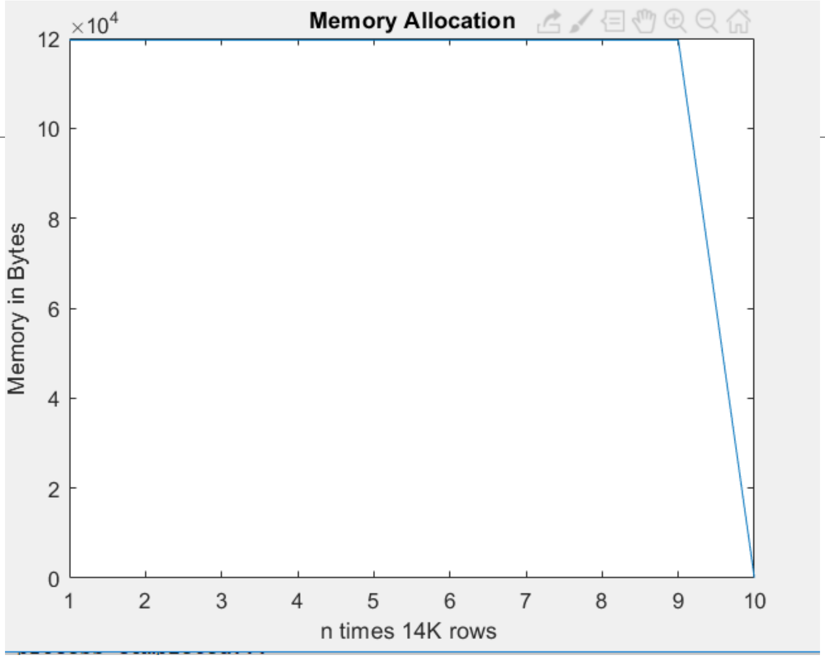


The model can run efficiently

- O(n) * cons(t)

Run time and memory





The new model (reconstruction), the right side has expensive memory at the outset due to Pre-allocation and delete all unused allocated after the model.

Advantages:

- Determine if the computing machine can accommodate large-scale dataset before running.
- Ensure that the program will run smoothly without interruption due to out of memory error.
- Run-time efficiency.
- Memory efficient after running.

Implications

- Inference of new entity with predicate based rely mostly on the positive triples in the knowledge graph.
- Positive latent factor values has more significance to the probability of inference rather than on negative latent factor outcome.
- The model is tested only Yago KG. Therefore, reliability using other dataset can not be ensured.
- The model is scalable to large dataset because of the efficient run time.

Recommendation

- More future research is suggested to establish the validity of the model.
- The model is useful for large-scale knowledge graph dataset implementation.
- Future researchers on using tensor factorisation to predict missing links on knowledge graph and other sparse data can use the model and the outcome of this study.

Thank you!