

# Ensemble NMF Tool - Manual

---

## Introduction

This document describes the Ensemble NMF command line tool, a C-based command line reference implementation of the *Ensemble NMF* clustering algorithm.

For further details regarding the algorithm, or to download the latest version of the software, please visit the *UCD Machine Learning Group* website<sup>1</sup>, or contact `derek.greene@ucd.ie`.

## Compilation Instructions

To compile the software requires:

- A working C++ compiler (*e.g.* GCC) and `make` tool.
- An implementation of the BLAS<sup>2</sup> (Basic Linear Algebra Subprograms) matrix toolkit. The ATLAS<sup>3</sup> (Automatically Tuned Linear Algebra Software) implementation is recommended.
- The CBLAS<sup>4</sup> interface to BLAS for C.

Unpack the source archive, then execute the `make` command in the source directory to build the `nmfens` binary.

## File Formats

The NMF Ensemble tool accepts a symmetric similarity matrix, which should be stored in the standard Matrix Market<sup>5</sup> format for sparse matrices. This is a coordinate-based format, containing a header followed by a line for each non-zero entry in the matrix.

For instance, the following file represents a  $4 \times 4$  symmetric real-valued matrix with 5 non-zero values:

```
%%MatrixMarket matrix coordinate real symmetric
4 4 5
1 1 1.000000
2 1 0.990000
2 2 1.000000
3 3 1.000000
4 4 1.000000
```

An optional list of data object identifiers can also be provided in a separate file, ending with the extension `.ids`. If the similarity matrix is stored in the file `sample.mtx`, then the identifier list should be stored in the file `sample.ids` in the same directory.

---

<sup>1</sup><http://mlg.ucd.ie/nmf>

<sup>2</sup><http://www.netlib.org/blas/>

<sup>3</sup><http://math-atlas.sourceforge.net/>

<sup>4</sup><http://www.netlib.org/blas/blast-forum/cblas.tgz>

<sup>5</sup><http://math.nist.gov/MatrixMarket/>

Each line in the file provides an object identifier for the corresponding row/column in the similarity matrix. The number of lines should correspond to the number of rows/columns in the similarity matrix. For instance, the a file corresponding to the matrix given previously might have the following content:

```
object_a
object_b
object_c
object_d
```

## Program Usage

The `nmfens` program is run from the command line, and requires at least one parameter – the path of the file containing the similarity matrix to cluster. For example, for the file `sample.mtx`:

```
./nmfens sample.mtx
```

A number of command line options can be used to customize the default parameters of the Ensemble NMF algorithm:

Option	Description	Default Value
<code>-m, --members</code>	Number of ensemble members to produce during the generation phase of the algorithm.	100
<code>--kmin</code>	Minimum value for number of basis vectors in each ensemble member.	5
<code>--kmax</code>	Maximum value for number of basis vectors in each ensemble member.	10
<code>--iters</code>	Maximum number of iterations for base clustering.	150
<code>--beta</code>	Parameter controlling the rate of convergence for symmetric NMF algorithm.	0.5
<code>--maxleaves</code>	Maximum number of leaf nodes in final soft hierarchy.	100

For example, to produce an ensemble clustering of the matrix contained in `sample.mtx`, where 100 ensemble members are generated and each member contains  $k \in [2, 5]$  basis vectors, we use the command:

```
./nmfens sample.mtx -m 100 --kmin 2 --kmax 5
```

Note that, for the experiments on the Collins protein interaction dataset described in our paper, we used the following command line parameters:

```
./nmfens collins.mtx -m 1000 --kmin 40 --kmax 60 --iters 150 --maxleaves 100
```